Optimisation d'une fonction de coût subjective sur un espace de filtres audio, et application à la modification émotionnelle de la parole

Julian Moreira

Stage du Master A.T.I.A.M.

Tuteur :	Jean-Julien Aucouturier
Centre d'accueil :	Riken Brain Science Institute (Wakoshi, Japon)
Date :	Avril 2010 - Juillet 2010

Remerciements

Je tiens à remercier tout d'abord mon maître de stage Jean-Julien Aucouturier, pour m'avoir fourni l'occasion de faire un stage sur un sujet aussi intéressant. Je le remercie également pour son encadrement, grâce auquel le stage a pu se dérouler dans les meilleures conditions. Il a su éveiller chez moi un intérêt constamment renouvelé. Je le remercie enfin pour tous ses conseils avisés, qui ont largement dépassé le simple cadre du travail.

Je remercie également Kazuo Okanoya (Okanoya-sensei), pour m'avoir accepté au sein de son laboratoire, ainsi que toute son équipe, dont l'accueil plus que chaleureux m'a permis de dépasser la barrière de la langue.

Je remercie Gentiane Venture et Isaku Kageyama, pour avoir permis cette collaboration extrêmement intéressante et enrichissante, et grâce à qui j'ai pu découvrir les joies de la capture de mouvement.

Enfin je remercie Céline Hamon, pour m'avoir accompagné tout au long de cette expérience, et sans qui ce stage aurait été beaucoup plus difficile.

Table des matières

Introduction

Que nous en soyons conscients ou non, notre voix transmet de nombreuses informations sur nos émotions. Le choix de nos mots et leur syntaxe bien sûr sont riches en information (voir par ex. [Constant *et al.*, 2008]). Mais même au delà des mots, le seul son de notre voix révèle souvent notre état émotionnel : joyeux, notre voix a de plus importantes variations de hauteur (pitch) que quand nous sommes tristes ; excités, notre débit est plus rapide [Scherer & Oshinsky, 1977]. Beaucoup d'études psychoacoustiques ont été menées pour établir des liens entre les paramètres acoustiques de la voix et les caractéristiques psychologiques qu'elle traduit, en utilisant des techniques d'analyse et synthèse de son [Scherer, 2003]. Beaucoup de travaux ont également été menés pour synthétiser des voix émotionnelles, par exemple pour rendre plus réalistes des personnages non-joueurs de jeux vidéos [Farner *et al.*, 2008], restaurer les voix dans de vieux films [Bertini *et al.*, n.d.], ou construire des guichets automatiques plus accueillants [Eide *et al.*, 2004] ; en revanche, on trouve peu de travaux sur le changement d'émotion sans changement de voix, i.e. en gardant l'impression d'entendre la même personne.

Le point de départ du stage est une expérience décrite dans [Aucouturier *et al.*, 2010], dont le but est justement de construire ce genre d'effets. Les auteurs montrent que de simples combinaisons d'effets audio-numériques standards, comme le vibrato, le trémolo ou variation de hauteur (pitch shifting), peuvent être paramétrées de façon à être :

- suffisamment rapides pour le temps-réel,
- suffisamment subtiles pour conserver le timbre du locuteur,
- et rester suffisamment efficaces pour provoquer une impression émotionnelle chez l'auditeur

Si la preuve de concept a été faite, les effets résultants présentent quelques désavantages :

- ils sont implémentés en hardware, ce qui les rend difficilement transportables
- les effets utilisés (effets temporels, spectraux, sur la hauteur) nécessitent tous un paramétrage indépendant, ce qui rend difficile d'explorer exhaustivement l'espace des paramètres
- les auteurs ont montré, en analysant trois types d'effet arbitraires, qu'on pouvait obtenir au moins trois émotions distinctes, mais il n'existe pas de technique pour synthétiser un effet visant une émotion prédéfinie

Le but de notre stage est de développer une solution à ces trois problèmes : un système capable de construire automatiquement des algorithmes de modification de la voix pour évoquer un grand nombre d'émotions, implémentables en logiciel

(software). Plus précisément, nous nous proposons de nous limiter à une seule classe d'effet audio-numérique : les filtres, qui sont à la fois facile à paramétrer, simples a implémenter, et potentiellement « émotionnellement efficaces ».

De façon intuitive, filtrer un signal consiste à accentuer ou atténuer l'énergie dans certaines zones de son spectre de fréquence. Les propriétés acoustiques de la voix émotionnelle ont rarement été décrites en terme de changements spectraux (e.g. un filtre passe-haut), car ces changements sont difficiles à produire physiologiquement (mais voir par ex. [Pittam Cynthia & Callan, 1990]). Toutefois, il est connu que certains filtres simples peuvent simuler avec succès des phénomènes vocaux associés à certaines émotions. Par exemple, l'excitation dun locuteur se traduit souvent par une plus forte énergie dans les hauteurs fréquences, rendant son timbre de voix plus « clair » et « tranchant » [Pittam Cynthia & Callan, 1990] - cela peut être simulé par un simple filtre passe-haut. À l'inverse, la tristesse est souvent associée à des voix décrites comme « sombres » et « mornes » – ce qu'un filtre passe-bas peut tout a fait évoquer. L'hypothèse de départ de notre travail est que nous pouvons généraliser ces effets au delà du simple passe-bas/passe-haut : il est possible que des filtres plus complexes (par exemple, un coupe-bande autour de 4000Hz) puissent évoquer une grande diversité d'émotions. Lenjeu de notre étude est de concevoir un système capable de les trouver, automatiquement.

D'un point de vue haut-niveau, notre problème consiste en fait à optimiser, sur l'espace des filtres audio, une fonction de coût mesurant l'adéquation entre un filtre et une émotion donnée (voir Figure 1). Cette fonction de coût est en fait subjective : elle ne résulte pas d'un algorithme que l'on puisse implémenter (comme par exemple, une fonction d'erreur de moindre carrés), mais d'un jugement psychologique nécessairement produit par un sujet humain. Chaque filtre considéré pendant l'optimisation doit être « testé en laboratoire » avant de pouvoir être inclus dans la recherche. Par exemple, pour déterminer le coût d'un filtre donné (ex. passe-bas, de fréquence de coupure 300Hz, gain +10dB), nous devons présenter plusieurs types de voix modifiées par ce même filtre à plusieurs sujets, et leur demander d'évaluer à quelle point ces voix évoquent e.g. l'émotion « tristesse ».

Cette situation est différente d'un apprentissage automatique classique, où les données dapprentissage sont fournies a priori, et l'on cherche à optimiser l'adéquation (souvent en terme probabilistes) entre un modèle et les données. Dans notre cas, l'algorithme d'optimisation décide quelles données doivent être recueillies (i.e. quels filtres doivent être testés sur sujet humains) pour faire avancer la recherche. L'algorithme précède la collecte de données.

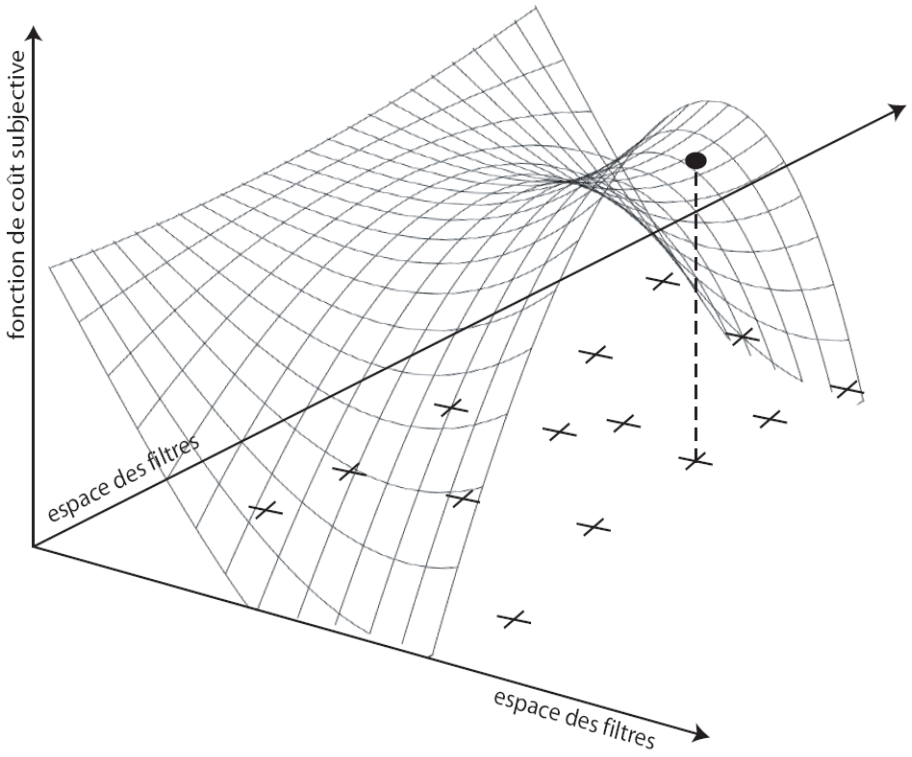


FIGURE 1 – *Exemple de fonction de coût implémentée sur l'espace des filtres.*

Ce rapport s'organise de la manière suivante : dans un premier temps, nous décrivons l'algorithme d'optimisation mis en place, et en particulier une méthode de paramétrisation de l'espace des filtres et une technique de recherche dans cet espace basée sur un échantillonnage par filtres tests. Dans une deuxième partie, nous décrivons une importante campagne de collecte de données visant à échantillonner les fonctions de coûts correspondant à 50 mots émotionnels – et comment nous avons conduit des tests d'écoute sur plus de 120 sujets. Enfin, nous présenterons le travail restant a réaliser pendant le dernier mois du stage (ce rapport correspond au troisième mois), qui comprend la validation des résultats de l'algorithme d'optimisation, par d'autres tests d'écoute.

1 Algorithme d'optimisation

Cette section présente la méthode choisie pour paramétrer l'espace des filtres, ainsi qu'une technique de recherche dans cet espace basé sur un échantillonnage par filtres tests. Dans un premier temps nous décrivons les principes généraux de ces méthodes, puis nous détaillons les algorithmes implémentés.

1.1 Paramétrer et créer des filtres optimaux

Le nombre limité de sujets humains – qui doivent, rappelons-le, juger nos filtres tests – entraine certaines contraintes concernant la paramétrisation de l'espace des filtres : tout d'abord, ces paramètres doivent être en nombre suffisamment réduit pour permettre à nos filtres tests de recouvrir au mieux cet espace (et nous éviter ainsi d'être confronté au « Fléau de la dimension », ou « Curse of dimensionality »). De plus, pour que la fonction de coût issue de notre exploration soit la plus juste possible, nous cherchons également à éviter des paramètres qui présenteraient des discontinuités importantes (nous pensons par exemple aux coefficients temporels des filtres, au voisinage de leurs pôles).
À ce titre, l'idée de diviser nos filtres tests en bande de fréquences (inspirée d'une méthode de personnalisation d'égaliseur proposée par [Sabin & Pardo, 2009b]), nous semble être un bon compromis : au nombre de quarante (voir Figure 2), ces bandes de fréquences présentent en plus cet avantage de couvrir l'espace des filtres de manière continue.

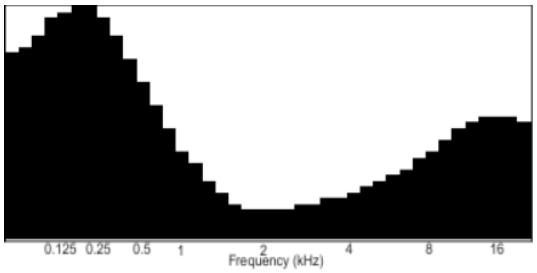


FIGURE 2 – *Exemple d'égaliseur. Le filtre est divisé en quarante bandes, et chaque bande possède son propre gain en décibels. Image tirée de [Sabin & Pardo, 2009a].*

A noter ici que les signaux audio sur lesquels seront appliqués ces filtres tests doivent être divisés en quarante bandes. Nous utilisons pour ça un banc de filtre ERB (Equivalent Rectangular Bandwidth, voir Figure 3) [Slaney, 1998]. Basé sur des mesures psychoacoustiques, ce système est conçu de manière à reproduire la perception du système auditif périphérique. Chaque filtre est un passe-bande, dont la largeur de bande est donnée à partir de sa fréquence centrale f_c, par la formule suivante [Slaney, 1993] :

$$BW_{ERB} = 24.7(0.00437 f_c + 1)$$

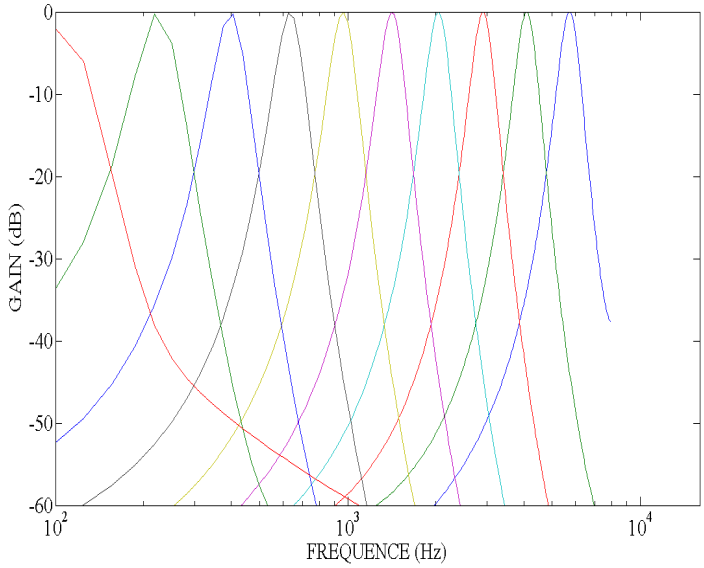


FIGURE 3 – *Exemple de 10 filtres ERB, échantillonnés à 16000 Hz.*

Pour couvrir au mieux cet espace des filtres, nous générons des filtres tests les plus différentes possibles. Les paramètres de ces courbes sont ensuite pondérées par des notes, attribuées par les sujets, et un filtre résultant est calculé à partir de ces données.
Quant à la manière de calculer ce filtre résultant, nous utilisons également inspirée de [Sabin & Pardo, 2009b], mais néanmoins différente : une fois les notes attribuées par les sujets, pour chacune des quarante bandes de fréquences, nous représentons nos résultats comme une distribution de notes sur un ensemble de gains. Nous calculons ensuite le barycentre de ces gains pondérés par les notes. Les détails de la méthode sont présentés en partie 1.2.2. Nous calculons donc en tout quarante barycentres, un par bande, et chacun d'eux sera un point de notre filtre résultant.

La différence de méthode avec le papier cité provient du fait que nous faisons noter nos filtres tests par plusieurs sujets, alors que les auteurs du papier cherchent justement à créer un égaliseur personnalisé. Dans un tel cas, chacune des notes a son importance, alors que nous cherchons au contraire à dégager des tendances, en s'affranchissant au maximum de la variation de sujet à sujet.
Imaginons par exemple, pour la bande 1, que toutes les notes obtenues soient nulles, sauf une note très élevée à 7dB. Dans le cas de [Sabin & Pardo, 2009b], l'utilisateur aura trouvé que ce gain à 7dB est le seul qui représente l'adjectif considéré. Il est donc important d'en prendre compte. En revanche dans notre cas,

cela signifiera que personne n'aura trouvé cette bande représentative de l'adjectif, exceptée une personne, à 7dB. Nous devons donc considérer que globalement, la bande 1 a été mal notée par les sujets, et nous tiendrons compte de ce critère pour la construction de notre filtre résultant.
Une deuxième raison nous a poussés à calculer le filtre résultant différemment : dans le papier cité, les auteurs obtiennent d'abord une « courbe de poids », qu'ils doivent normaliser entre -1 et 1, puis multiplier par 20 pour enfin avoir un égaliseur entre -20dB et 20dB. Le fait de normaliser revient à ne jamais tenir compte des valeurs des gains des « filtres test », juste de leur allure générale. Notre objectif étant de construire un filtre directement applicable, et non pas un effet ajustable avec un curseur, notre calcul de barycentre nous permet de nous affranchir de cette considération.

A noter enfin que tout au long de cette méthode, nous faisons une hypothèse importante : toutes les bandes de fréquences sont indépendantes les unes des autres. Cette hypothèse peut sembler assez lourde au premier abord. En effet, si par exemple l'adjectif « joyeux » se caractérise par une opposition de gains entre la bande 10 et la bande 30 (bande 10 élevée et bande 30 basse, ou l'inverse), alors les notes distribuées par les sujets seront élevées dans les deux cas. Seulement notre algorithme sera uniquement capable de voir que la bande 10 a des notes élevées à la fois pour un gain élevé et pour un gain bas. Il conclura donc que la bande 10 ne caractérise nullement l'adjectif « joyeux », et cette bande sera nulle dans le filtre résultant. Idem pour la bande 30, et le filtre résultant sera alors complètement nul. Finalement, cela revient à dire qu'avec cette hypothèse nous supposons que pour chaque bande, la distribution des notes sur l'ensemble des gains est unimodale.
Cependant, l'hypothèse est déjà faite dans [Sabin & Pardo, 2009b]. Aux vues de leur résultats positifs, et suite à une conversation privée avec les auteurs, il nous semble raisonnable de la conserver.

1.2 Algorithme d'apprentissage

L'algorithme d'apprentissage se divise en deux parties : la première génère les filtres tests, et la seconde construit les filtres résultants une fois les données récoltées. Tous les programmes ont été implémentés en Matlab.

1.2.1 Génération des filtres tests

Afin de sélectionner les filtres tests les plus variés possibles, notre algorithme de génération comporte lui-même de deux étapes : la première génère mille courbes aléatoirement, dont seulement un certain nombre sera sélectionnées au cours de

l'étape suivante. On estime que trente courbes par adjectif suffisent à construire le filtre résultant. Ces trente courbes seront donc notées, pour chaque adjectif, au cours de l'expérience.

Pour créer une des mille courbes, nous choisissons de faire une combinaison linéaire d'un nombre aléatoire de distributions gaussiennes, compris entre deux et huit, dont les paramètres (moyenne, écart-type, amplitude maximale et offset) sont eux aussi aléatoires. Dans tous les cas, nous prenons garde que l'amplitude des courbes ne dépasse jamais 10 en valeur absolue, pour nous assurer que notre filtre sera bien compris entre -10dB et 10dB (dans le cas contraire nous normalisons la gaussienne en la divisant par le maximum de sa valeur absolue). Cette bande de gain, nous l'avons choisie arbitrairement, pour que les filtres tests générés, une fois appliqués à la voix, produisent des stimuli réalistes, i.e. imaginables dans un contexte de discussion réelle.
Notons ici que la décision de créer nos filtres tests avec des distributions gaussiennes, plutôt que de générer juste quarante points aléatoirement, ou d'utiliser par exemple des portes, provient du fait que nous voulons avoir des effets physiquement réalisables par la voix, en théorie au moins, c'est-à-dire sans passage abrupt d'une fréquence à la suivante (par exemple un gain de 10dB à 1000 Hz et un autre de -10dB à 1001 Hz). Ce choix n'est pas sans conséquences, puisqu'il entraine une concentration de points plus importante vers l'amplitude 0 que vers les amplitudes extrêmes (voir Figure 4). Il en sera bien sûr de même pour les trente courbes sélectionnées, et sur cet ensemble de courbes, chaque bande de fréquences sera donc plus souvent notée pour une valeur proche de 0dB que pour les valeurs -10dB et 10dB. Nous verrons dans la partie suivante l'impact que cette hétérogénéité dans la répartition des gains peut avoir sur le calcul du barycentre, et la manière dont nous y remédions.

Les trente courbes choisies parmi ces mille ne sont en réalité que vingt-deux. En effet, nous décidons que nos huit premiers filtres seront des filtres standards, couramment utilisés et donc couramment entendus : un passe-bas (fréquence de coupure $f_c = 1000Hz$), un passe-haut ($f_c = 10000Hz$), un coupe-bas ($f_c = 1000Hz$), un coupe-haut ($f_c = 10000Hz$), deux passes-bandes ($f_{bp} \in [2000Hz\ 6000Hz]$, $f_{bp} \in [3500Hz\ 4500Hz]$) et deux coupes-bandes ($f_{bc} \in [2000Hz\ 6000Hz]$, $f_{bc} \in [3500Hz\ 4500Hz]$).

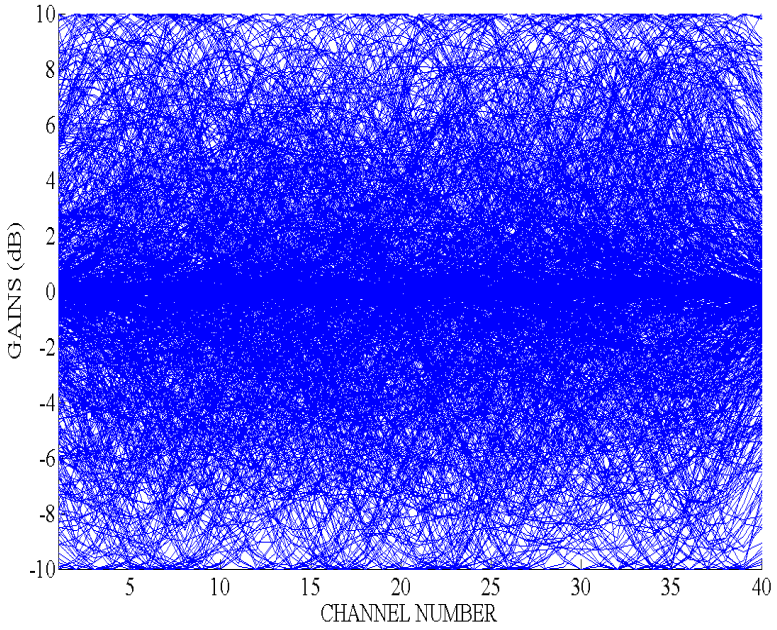


FIGURE 4 – *Exemple de 1000 courbes générées aléatoirement, puis superposées. On discerne bien une bande plus épaisse aux alentours de 0dB, alors que les extrémités sont plus clairsemées.*

Enfin, l'algorithme itératif suivant est répété jusqu'à l'obtention complète des vingt-deux courbes encore à sélectionner :
- on prend une à une toutes les courbes générées aléatoirement (à la première itération il y en a mille donc) ;
- pour chacune d'entres elles et pour chaque bande de fréquence on calcule la déviation standard du gain avec les gains des courbes déjà sélectionnées (on a donc à chaque tour quarante déviations standards) ;
- on fait la moyenne de ces déviations standards ;
- on choisit la courbe dont la moyenne est la plus grande ;
- on la retire de la liste des courbes générées aléatoirement.

Cette méthode nous permet donc d'obtenir trente courbes, a priori les plus différentes les une des autres, en tout cas qui prennent ensemble un maximum de place dans l'espace des gains. La Figure 5 permet d'avoir un aperçu du genre de courbes qu'on peut obtenir.

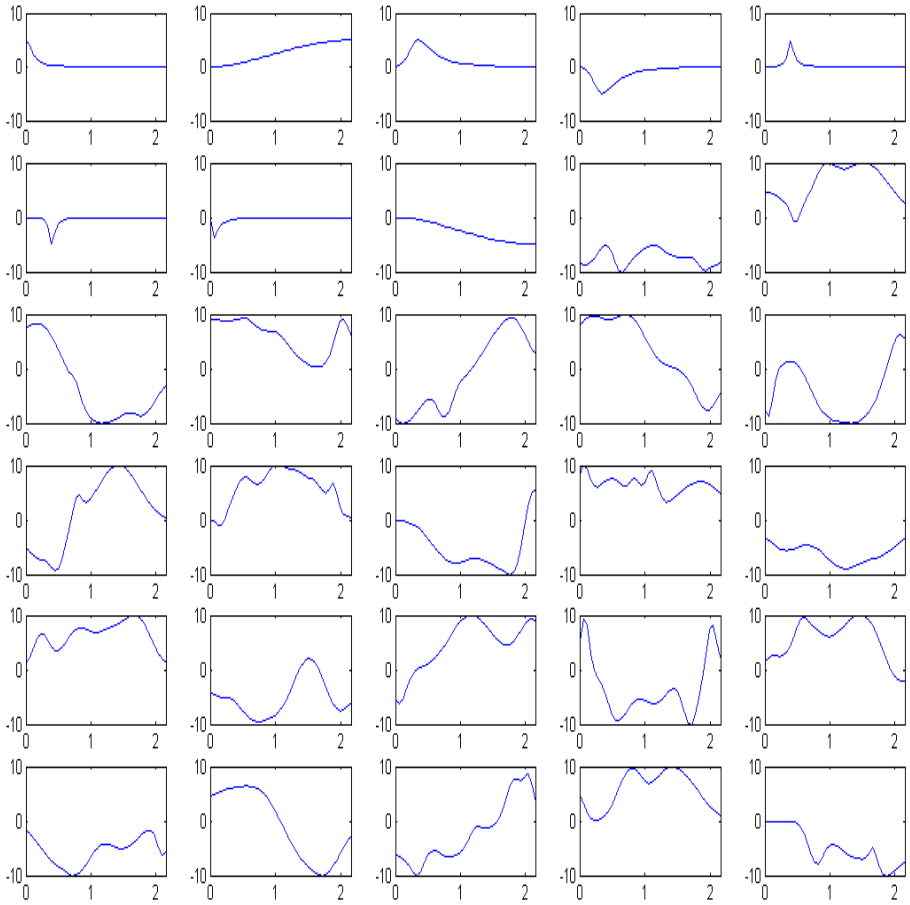


FIGURE 5 – *Trente filtres tests. Les huit premiers sont les filtres usuels, les autres les courbes sélectionnées parmi mille. En abscisse de chaque courbe la fréquence en kHz, en ordonnée le gain en dB.*

1.2.2 Construction du filtre résultant

Pour chaque bande de fréquences, on affiche les notes de toutes les courbes en fonction de leur gains. La Figure 6) est un exemple de fonction de répartition qu'on peut obtenir.

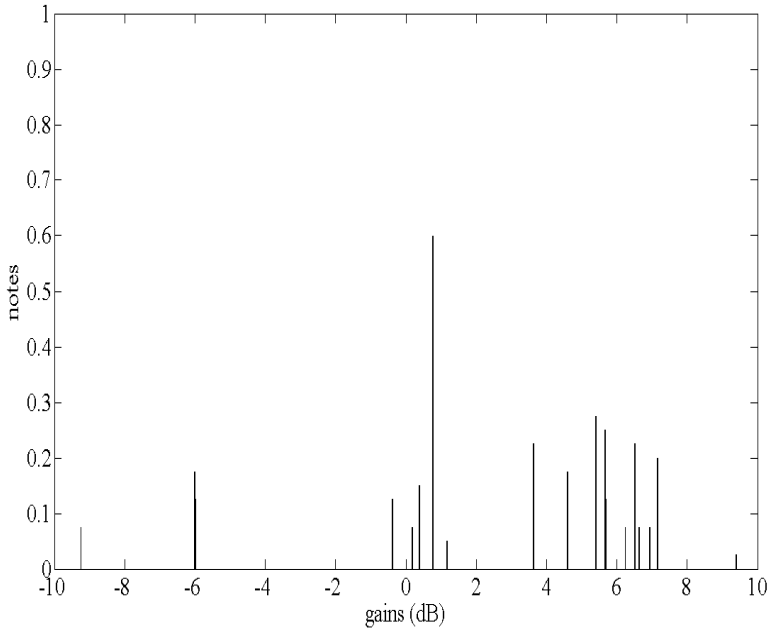


FIGURE 6 – *Exemple de fonction de répartition des notes sur les gains, pour une bande de fréquence donnée.*

Le barycentre de cette répartition constituera un point de notre filtre final. Le barycentre est définit comme étant la somme de tous les gains multipliés par leur note, et le tout divisé par la somme des notes.
Cependant, c'est ici que le problème évoqué dans la partie 1.2.1 refait surface : chaque courbe étant plus souvent notée pour des gains proches de 0dB que pour les extrémités, le barycentre sera constamment biaisé, et « ramené » vers 0dB. Pour palier à ce problème, nous implémentons une fonction calculant des points d'interpolation sur l'ensemble de la fonction de répartition, de manière à ce qu'elle soit uniformément répartie. La Figure 7 illustre la manière dont elle procède : elle divise l'axe des gains en segment égaux, et pour chacun de ces segments, elle regarde si des gains notés se trouvent à l'intérieur. Si c'est le cas, elle calcule la moyenne de ces notes, et trace une droite entre cette moyenne et la moyenne précédemment calculée. Tous les segments vides entre ces deux points se verront attribuer une note sur cette droite.

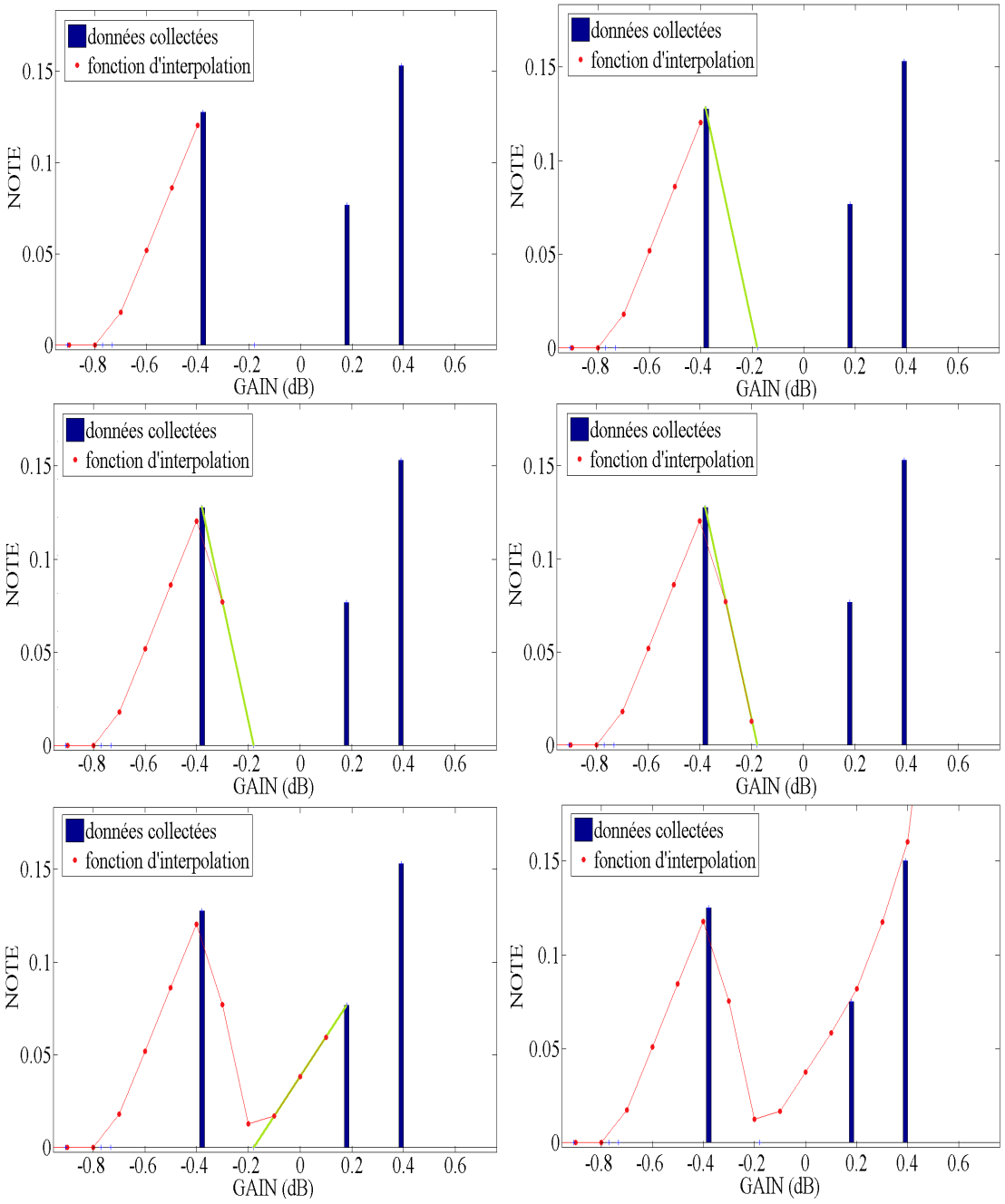


FIGURE 7 – *Principe de la fonction d'interpolation.*

Suite à cette opération, un barycentre sans biais est calculé à partir de la fonction de répartition interpolée (Figures 8-A, B et C). Il constitue dans le filtre résultant la valeur de gain de la bande de fréquence considérée.

Les barycentres des quarante bandes de fréquences calculés nous permettent donc d'obtenir le filtre résultant (Figure 8-D). Ce filtre est censé représenté au mieux, en fonction des notes attribuées aux courbes tests, l'adjectif considéré.

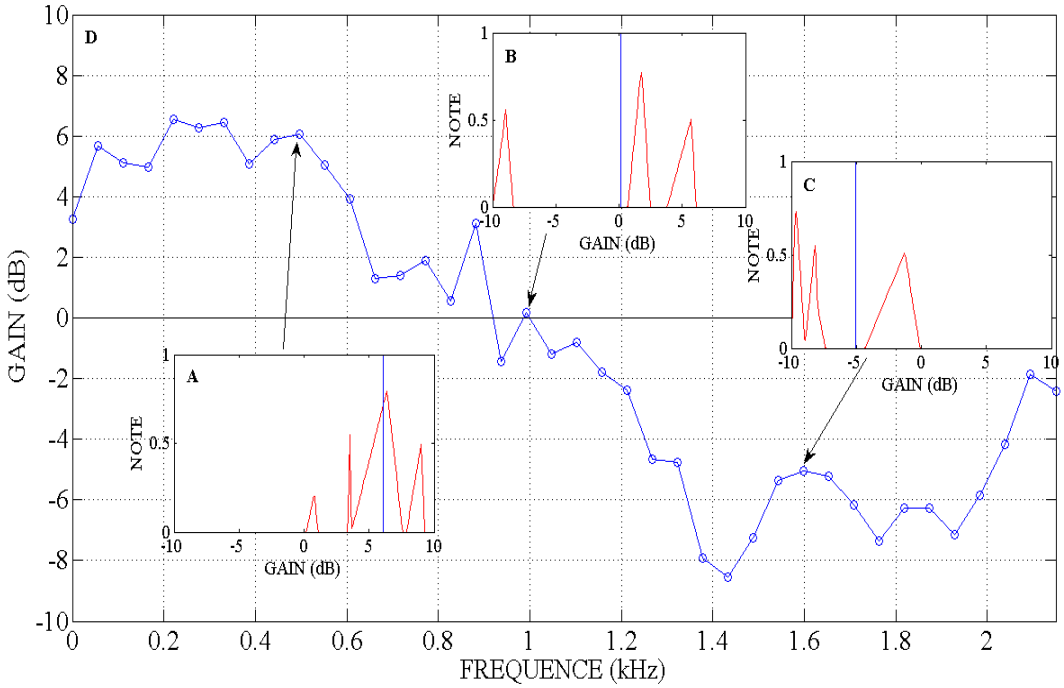


FIGURE 8 – *Exemple de filtre résultant, obtenu par calcul des barycentres. Dans les Figures A, B et C, la ligne bleue correspond au barycentre calculé.*

Afin de confirmer l'efficacité de notre méthode, nous pouvons dès à présent créer nous même un ensemble de notes, dont les valeurs dépendront des gains des filtres tests. Par exemple, La Figure 8-D est construite de cette manière. Ici, les notes ont été créées de la manière suivante : plus les gains des bandes de fréquences 10 et 30 sont respectivement élevées et basses pour un filtre test, plus sa note est bonne. Le filtre résultant révèle bien cette tendance, pour 30 filtres tests notés (un nombre assez petit), avec cette caractéristique que les bandes proches de la dixième et de la trentième sont aussi influencées par ce critère. Ce phénomène vient du fait que les filtres tests étant générés par des distributions gaussiennes, à chaque fois que la dixième bande sera élevée dans un filtre test, les bandes voisines le seront aussi.

2 Génération des fonctions de coût

Cette section détaille une application de l'algorithme présenté précédemment : les filtres tests sont appliqués à des signaux de voix, et les notes qui leur sont attribuées par les sujets concernent l'émotion qu'elles traduisent. Nous présentons d'abord une méthode pour choisir un panel de mots émotionnels à attribuer aux courbes, puis l'expérience mise en place pour collecter les données.

2.1 Choisir des émotions

Concrètement, les émotions que nous ressentons sont associées dans la vie courante à toutes sortes de mots plus ou moins affectifs : tristesse, joyeux, cimetière, extrêmement, surprendre, etc. Pour affecter des émotions à des filtres, nous devons donc choisir les mots qui représenteront au mieux ces émotions. Un effet « joyeux » étant bien plus parlant pour une voix qu'un effet « cimetière », ou qu'un effet « surprendre », nous nous bornons à ne prendre que des adjectifs.
Notre objectif étant de tester l'éventail d'émotions qu'on peut représenter avec un simple filtre, nous nous appuyons de plus sur le papier de Bradley M.M. et Lang P.J. [Bradley & Lang, 1999], qui fournit un corpus important de mots notés sur une échelle émotionnelle à trois axes, couramment utilisée aujourd'hui, l'échelle « arousal - valence - dominance »(chaque axe allant de 0 à 10, voir Figure 9). Par ailleurs, la méthode de notation utilisée est elle aussi largement éprouvée, le « mannequin d'auto-évaluation » (« the self-assessment manikin », voir [Bradley & Lang, 1994]).

Étant donné le nombre important de mots (il y en a 1034, qui ne sont pas tous des adjectifs par ailleurs), nous ne pouvons pas construire de filtre résultant pour chacun d'entre eux (le nombre de sujets à trouver serait trop important). Nous voulons donc un nombre plus réduit de mots, mais qui pave malgré tout l'espace des émotions à peu près uniformément. C'est la raison pour laquelle nous choisissons d'implémenter un algorithme k-means (voir [Candillier, 2006]), qui sélectionnera pour nous cinquante clusters (ce nombre a été choisi en fonction du nombre de sujets que nous pensons réussir à trouver), et dans lesquels on pourra ensuite trouver un adjectif.

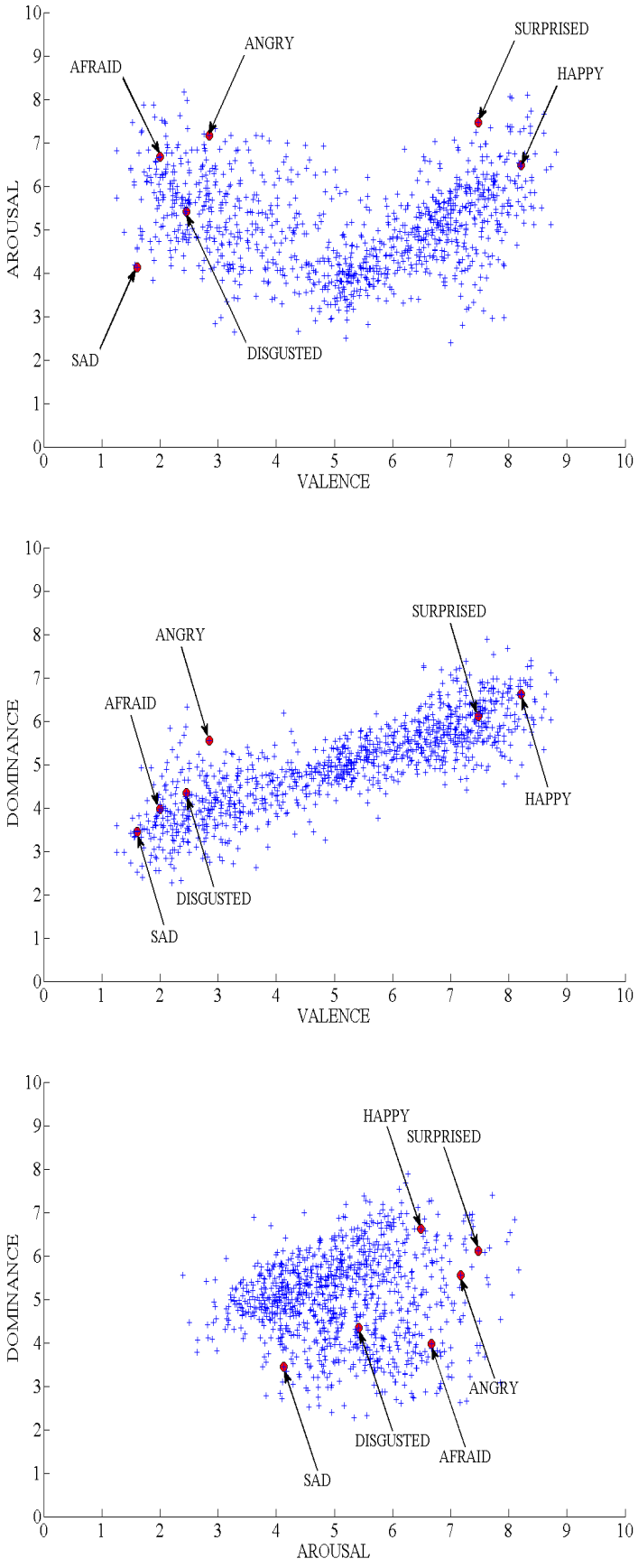


FIGURE 9 – *Représentation des 1034 mots de [Bradley & Lang, 1999] dans l'espace « valence-arousal-dominance ». En rouge, les six ajectifs correspondants aux émotions dites « universelles » : heureux (happy), apeuré (afraid), triste (sad), dégoûté (disgusted), en colère (angry) et surpris (surprised).*

Notre algorithme k-means est constitué de trois étapes simples :
- à l'initialisation d'abord, cinquante mots sont choisis aléatoirement. Ils sont les centroïdes de nos clusters ;
et tant que les cinquante clusters diffèrent d'une itération à la suivante :
- pour chaque mot de l'espace, on calcule sa distance euclidienne à chaque centroïde, et on le définit comme appartenant au cluster du centroïde le plus proche ;
- on calcule pour chaque cluster le point dont les coordonnées sont égales à la moyenne des coordonnées de chaque point du cluster, et on définit le nouveau centroïde par le mot le plus proche de ce point.
La Figure 10 illustre cet algorithme, fait à titre d'exemple avec deux clusters.

L'algorithme k-means présente l'inconvénient de converger vers des solutions optimales locales, et pour cette raison il est courant d'itérer l'algorithme plusieurs fois, en changeant aléatoirement les centroïdes initiaux, puis de choisir les clusters en fonction d'un critère de qualité interne. Dans notre cas cependant le nombre de clusters que nous cherchons à former est trop important par rapport au nombre de points total (50 clusters pour 1034 mots). Pour cette raison, la convergence des solutions est presque à chaque fois différente (à peu de choses près) d'une itération de l'algorithme à l'autre, et ce quelque soit l'initialisation. Comme nous cherchons ici simplement à choisir un adjectif, au sein de chaque cluster, de telle manière que l'ensemble choisi pave uniformément l'espace (ce qui est presque tout le le temps le cas avec cinquante clusters), nous ne nous soucions pas davantage de cet éventuel problème.
Nous choisisons pour finir nos adjectifs à la main, un dans chaque cluster donc, de telle manière qu'ils soient le plus applicable possible à la voix. L'ensemble des mots choisis sont visibles sur la Figure 11, et dans la Table 1.

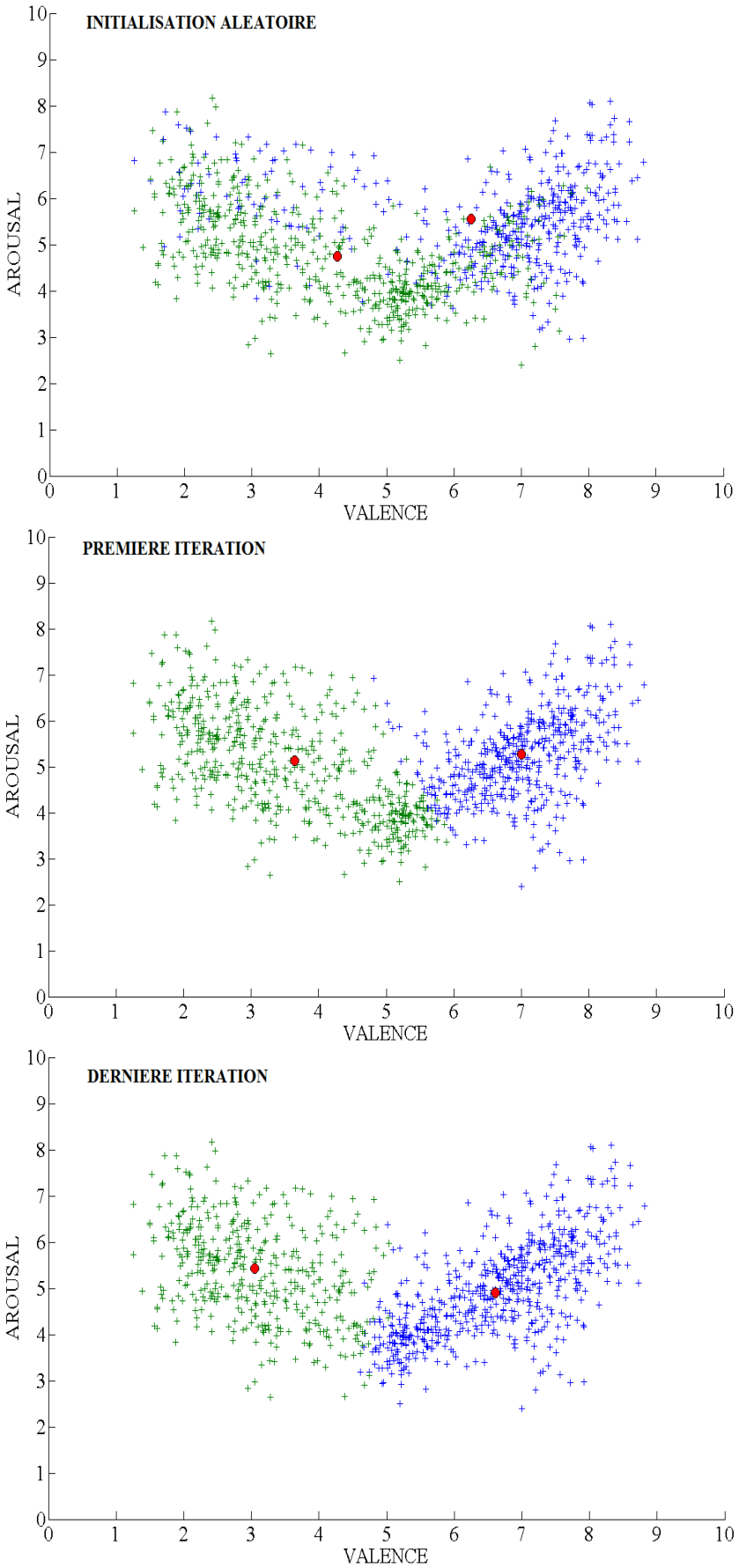


FIGURE 10 – *Les deux premières et la dernière itérations de l'algorithme k-means effectué sur l'ensemble des mots, pour deux clusters (à titre d'exemple). L'algorithme est calculé sur les données en trois dimensions, et représenté ici en deux dimensions par souci de lisibilité. En rouge les centroïdes.*

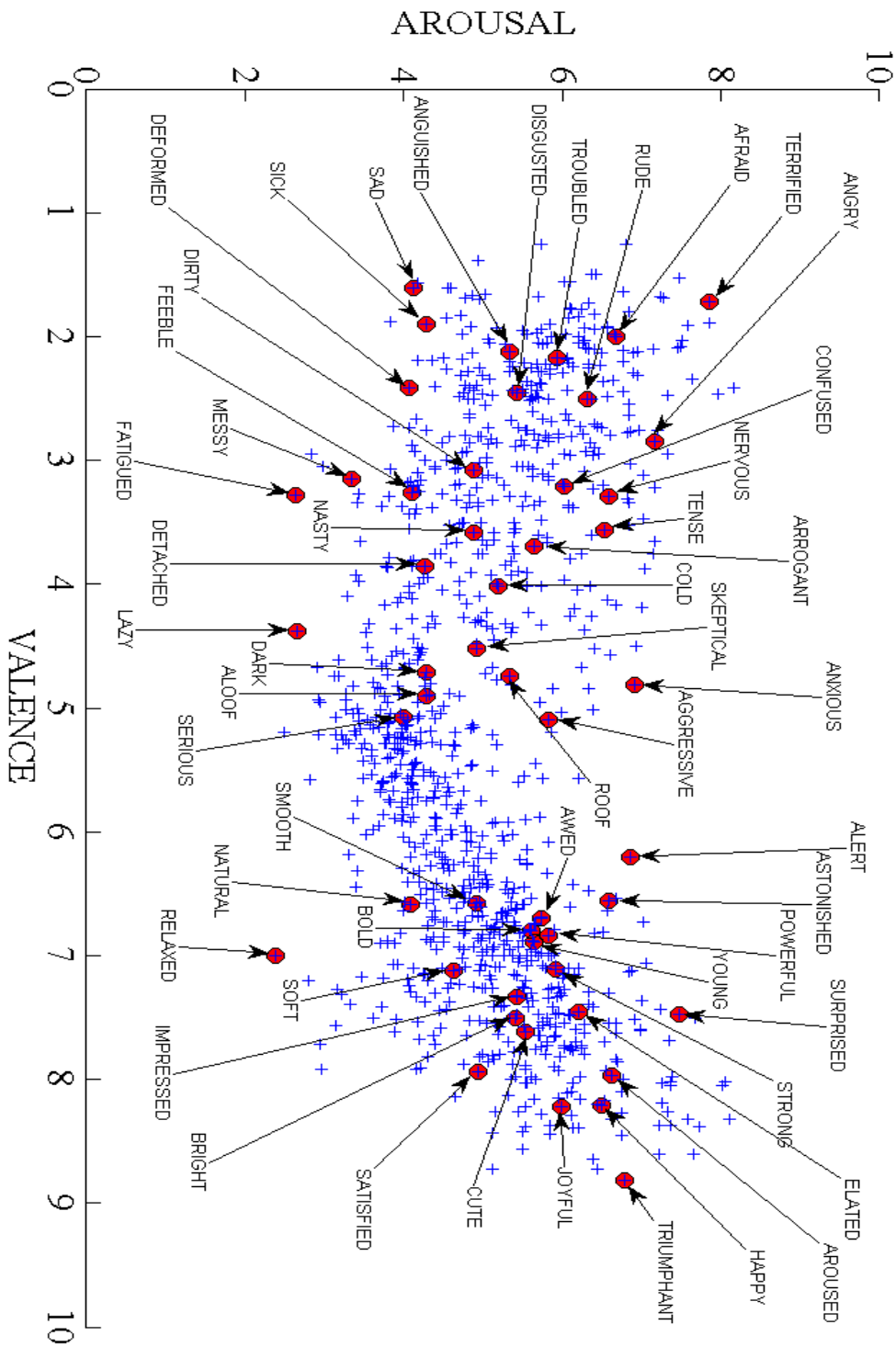


FIGURE 11 – *Représentation des 1034 mots affectifs (en 2D, sur le plan valence-arousal), avec en rouge les adjectifs choisis.*

adjectif	valence	arousal	dominance	adjectif	valence	arousal	dominance
afraid	2	6.67	3.98	impressed	7.33	5.42	5.51
aggressive	5.1	5.83	5.59	joyful	8.22	5.98	6.6
alert	6.2	6.85	5.96	lazy	4.38	2.65	4.07
aloof	4.9	4.28	4.69	messy	3.15	3.34	4.75
angry	2.85	7.17	5.55	nasty	3.58	4.89	5
anguished	2.12	5.33	3.45	natural	6.59	4.09	5.57
anxious	4.81	6.92	5.33	nervous	3.29	6.59	3.56
aroused	7.97	6.63	6.14	powerful	6.84	5.83	7.19
arrogant	3.69	5.65	5.14	relaxed	7	2.39	5.55
astonished	6.56	6.58	5.16	rough	4.74	5.33	4.81
awed	6.7	5.74	5.3	rude	2.5	6.31	4.91
bold	6.8	5.6	6.67	sad	1.61	4.13	3.45
bright	7.5	5.4	6.34	satisfied	7.94	4.94	6.14
cold	4.02	5.19	4.69	serious	5.08	4	5.12
confused	3.21	6.03	4.24	sick	1.9	4.29	3.04
cute	7.62	5.53	4.86	skeptical	4.52	4.91	4.5
dark	4.71	4.28	4.84	smooth	6.58	4.91	5.09
deformed	2.41	4.07	3.95	soft	7.12	4.63	6
detached	3.86	4.26	3.63	strong	7.11	5.92	6.92
dirty	3.08	4.88	4.7	surprised	7.47	7.47	6.11
disgusted	2.45	5.42	4.34	tense	3.56	6.53	5.22
elated	7.45	6.21	5.53	terrified	1.72	7.86	3.08
fatigued	3.28	2.64	3.78	triumphant	8.82	6.78	6.95
feeble	3.26	4.1	2.71	troubled	2.17	5.94	3.91
happy	8.21	6.49	6.63	young	6.89	5.64	5.3

TABLE 1 – *Liste des adjectifs sélectionnés à l'aide de l'algortihme k-means*

2.2 Collecte des données d'apprentissage

La collecte des données d'apprentissage est une expérience à part entière, nécessitant, en plus de l'organisation concrète du test, de nombreux « à côtés » : création des stimuli, implémentation de pages web, d'une base de données, et de programmes pour la gérer.

2.2.1 Stimuli

Nous choisissons pour créer nos stimuli des phrases simples, courtes, en anglais, et certifiées comme étant neutres par [Russ *et al.*, 2008]. Elles sont au nombre de 14, et sont visibles en annexe. Ces phrases ont été enregistrées par 6 personnes

différentes (3 femmes), tous natifs d'un pays anglophone, et au moyen d'un enregistreur Sony PCM D50. Pendant la phase d'enregistrement, nous avons demandé à chaque personne de lire à voix haute les phrases de la manière la plus neutre possible (i.e. sans aucune émotion), deux fois de suite, puis une dernière fois d'une manière cette fois plus « émotionnelle », avec l'émotion de leur choix (juste à titre de comparaison, nous n'utilisons pas ce dernier jeu de phrases lors de l'expérience). Nous avons ensuite sélectionné parmi les deux premiers jeux les phrases qui nous ont semblé être les plus neutres. Ces phrases neutres (au nombre de 84) constituent la première partie de nos stimuli.

Pour chacune de nos 1108 filtres tests, nous voulons créer 5 stimuli différents (la deuxième partie des stimuli), de telle sorte que les phrases et les locuteurs soient à chaque fois différents. Pour cela nous implémentons un programme Matlab, qui choisit 5 locuteurs et 5 phrases de la manière suivante : comme il y a en tout 6 locuteurs, nous faisons en sorte que celui qui manque soit choisi pour le filtre test suivant. Les cinq choisis sont permutés aléatoirement. La parité homme/femme est également respectée : 3 femmes et 2 hommes pour un filtre test, et l'inverse pour la suivante. De même pour les 14 phrases, avant de piocher une deuxième fois la même, nous sélectionnons toutes celles qui n'ont pas déjà été choisies, dans un ordre aléatoire.
De cette manière on s'assure que l'association locuteur+phrase est aléatoire, en respectant quand même un certain nombre de règles.

Au final, on a 7500 stimuli différents avec ces filtres tests et filtres usuels, c'est à dire 150 filtres tests/filtres usuels notés par mots, et 50 mots en tout. Nous devons rajouter à ces stimuli les 84 premiers, donc en tout 7584. Enfin, ces stimuli sont normalisés avec Audacity. Toutes les données concernant l'expérience, stimuli y compris, sont stockés dans une base de données implémentée en MySQL.

2.2.2 Participants

Pour le moment, 113 sujets ont participé (64 femmes, 49 hommes), d'une moyenne d'âge de 23 ans. La majorité des sujets étaient des étudiants de licence dans une université américaine implantée à Tokyo. L'anglais n'était pas nécessairement leur langue maternelle, mais ils le parlaient tous plus que convenablement.

2.2.3 Déroulement de l'expérience

Toute l'expérience se déroule via Internet, sur des pages web développées en PHP (visibles en annexe). Le sujet doit d'abord rentrer quelques informations le concernant, régler le volume sonore, et lire une page de tutoriel (toutes ces pages

sont visibles en annexe). Il se voit ensuite présenter une sélection de 50 paires de stimuli, piochées aléatoirement parmi les paires de stimuli non-notées. Chaque paire est composée d'un stimulus modifiée avec un de nos filtres tests, et le stimulus « neutre » (i.e. sans modification) qui lui correspond, c'est à dire avec la même phrase et le même locuteur. Le sujet doit noter quel stimulus correspond mieux au mot émotionnel affiché (voir Figure 12). Le sujet note une fois chaque mot, présentés dans un ordre aléatoire.

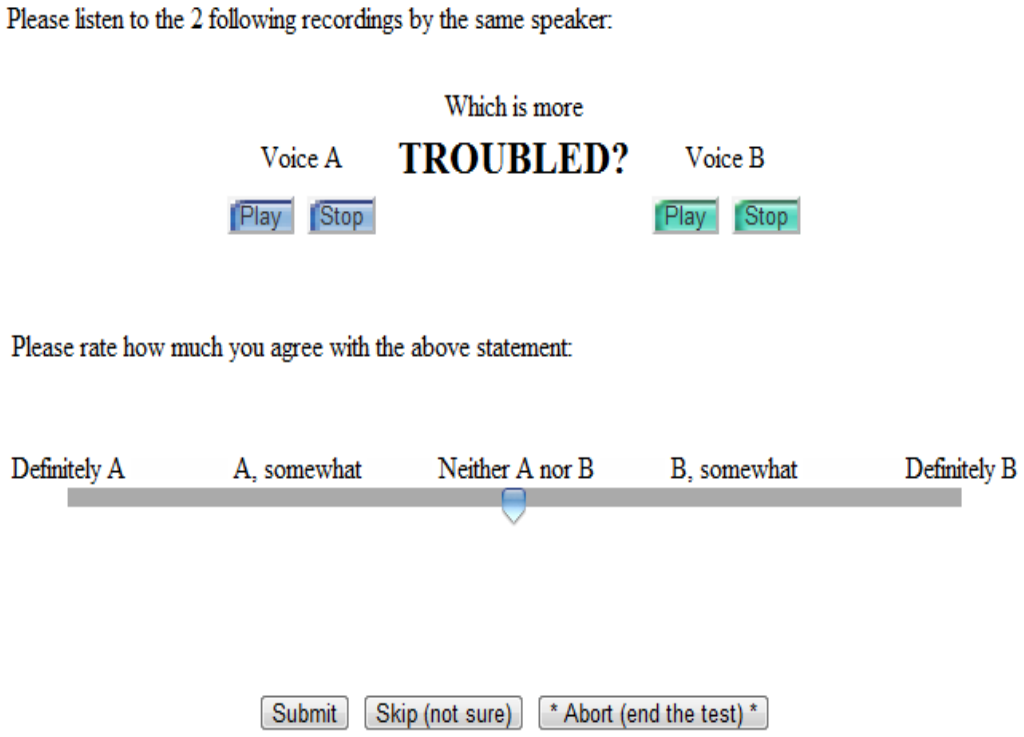


FIGURE 12 – *Page web présentée aux sujets, ici avec le mot « troubled » (troublé) à noter. La voix A correspond toujours au stimulus neutre, et la voix B correspond toujours au stimulus modifié par un filtre test.*

Le sujet a tout le temps qu'il veut pour répondre, mais une fois la réponse soumise, il ne peut plus revenir en arrière.

2.3 Informations diverses

L'expérience s'est déroulée à la Temple University of Japan, depuis le 15 juin 2010 jusqu'à aujourd'hui (le 26 juin 2010). Les sujets qui ont participé ont été réunis dans deux salles informatiques de l'université, et ont bénéficié d'un environnement calme tout au long de la procédure. Ils étaient de plus tous munis d'écouteurs. L'expérience s'est déroulée sur les ordinateurs de l'université. Elle a pris en moyenne 15 minutes à chaque sujet. Les sujets n'ont pas été rémunérés, mais ont été conviés à participer à un tirage au sort dont les gagnants remporteront 5000 yens (environ 40 euros).

3 Perspectives pour le dernier mois de stage

En plus de la fin de l'expérience (dépouillement des données et validation), ce dernier mois de stage sera pour nous l'occasion de poursuivre un autre projet entamé, sans lien avec le titre de ce rapport, autour du taiko, un instrument traditionnel japonais.

3.1 Dépouillement et validation des données collectées

Une fois les données entièrement récoltées (nous visons 150 sujets au total), nous pourrons construire nos filtres résultants, à l'aide des algorithmes décrits dans la partie 1. Cependant, même si ces filtres reflètent les notes attribuées par les sujets, ils ne seront pas forcément représentatif des mots émotionnels choisis. Pour s'en assurer, nous devons monter une deuxième expérience, de validation cette fois, durant laquelle des nouveaux sujets devront écouter, puis juger l'efficacité émotionnelle de stimuli modifiés par nos filtres résultants. Un des challenges de cette expérience résidera dans la mise en place de son protocole : en effet, elle devra respecter des normes psychologiques strictes, afin d'être acceptée par l'ensemble de la communauté scientifique. Nous prévoyons à ce titre une interaction avec des psychologues.
L'enjeu de cette expérience sera double : elle devra d'une part valider ou non nos filtres, et d'autre part mesurer l'étendue d'émotions que nous sommes capables de modéliser avec cet outil simple. Dans le cas de résultats probants, ces travaux pourraient mener à la rédaction d'une publication.

3.2 Étude du lien entre geste et son dans un instrument traditionnel japonais : le taiko

Le taiko est un instrument traditionnel japonais dont le nom pourrait se traduire par « gros tambour » (voir Figure 13). Instrument utilisé à l'origine pour accompagner les cérémonies bouddhistes, puis les représentations de théatre nô et kabuki, son emploi c'est largement diversifié depuis, allant même jusqu'à être synthétisé dans une musique de Kanye West, chanteur de hip-hop et de rap américain. C'est suite à une demande d'Isaku Kageyama, joueur professionnel de taiko, qu'une collaboration est née entre notre laboratoire et le laboratoire de robotique Gentiane Venture, de la Tokyo University of Agriculture and Technology. L'objectif, établi lors d'une réunion, est de comprendre les mouvements du joueur, afin d'en déduire des informations sur la nature des sons produits. A terme, ce projet pourrait mener à la création d'un système « d'air taiko », grâce auquel il serait possible de jouer du taiko sans instrument. Pour le moment, une séance de capture de mouvements a déjà eu lieu (associée à une capture de son, voir Figures 14 à 21), nous permettant de collecter les premières mesures, et de nous faire une idée de la variété de gestes et de sons possibles avec cet instrument. Nous prévoyons sur le mois qui arrive d'entamer le dépouillement des mesures, et d'étudier la corrélation entre mouvement et son.

FIGURE 13 – *Le taiko, un instrument traditionnel japonais.*

FIGURE 14 – *Salle de capture de mouvements.*

FIGURE 15 – *Isaku Kageyama, vêtu pour la capture de mouvement.*

FIGURE 16 – *Mise en place de la capture de son.*

FIGURE 17 – *Calibrage du logiciel de capture de mouvement (position en « T »).*

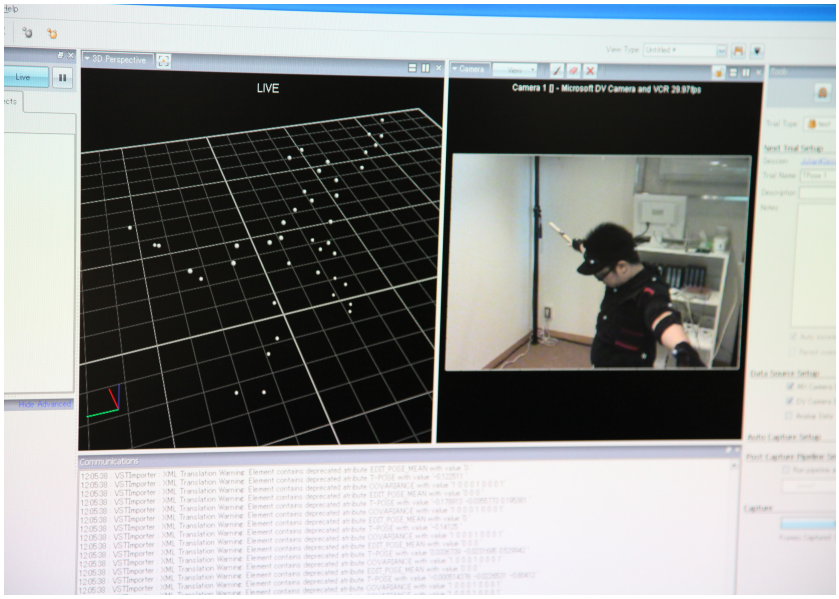


FIGURE 18 – *Le logiciel de capture de mouvement.*

FIGURE 19 – *La capture de mouvement s'est également faite par gyroscope.*

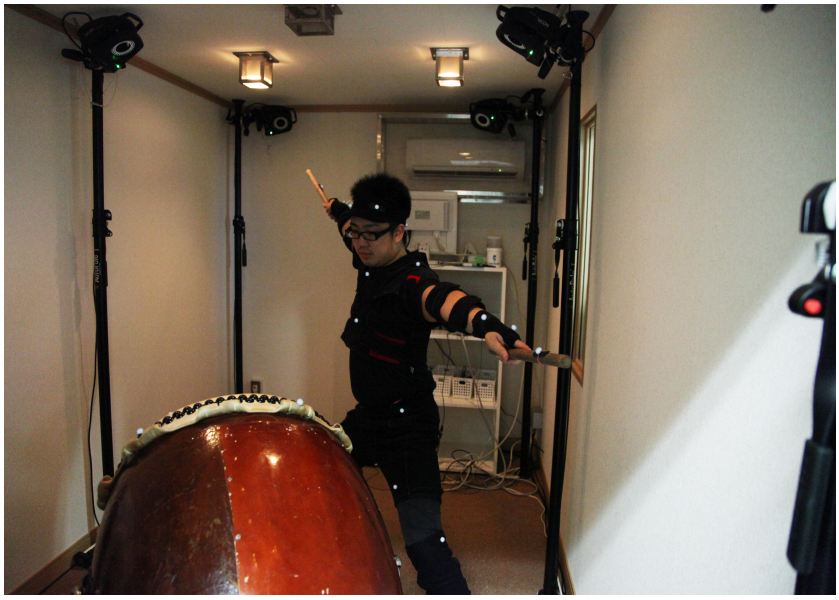


FIGURE 20 – *Pose traditionnelle prise pendant le jeu.*

FIGURE 21 – *Séance d'enregistrement.*

Conclusion

Dans ce rapport, nous avons proposé une méthode d'optimisation d'une fonction de coût subjective de l'espace des filtres. Dans l'application qu'on en fait, le système développé est capable de construire des filtres, qui, appliqués à la voix, peuvent potentiellement évoquer un grand nombre d'émotions. Une importante collecte de données a été organisée, visant à « nourrir » notre algorithme d'apprentissage, i.e. à noter des filtres tests qui échantillonnent notre espace de filtres. Si notre algorithme fonctionne, nous devons encore valider l'efficacité émotionnelle des filtres construits. Une expérience à ce propos est à venir très prochainement.

D'un point de vue plus personnel, ce stage aura été extrêmement enrichissante, et ce sur tous les plans. Professionnellement tout d'abord, puisqu'il m'aura permis d'étendre mes connaissances sur de nombreux sujets : psychologie, informatique (apprentissage de PHP, MySQL, approfondissement en Matlab), apprentissage, organisation d'une expérience, etc. J'ai également eu l'occasion de rédiger des revues de publications (sous la supervision de Jean-Julien Aucouturier) soumises pour la prochaine conférence d'ISMIR. Ce fut l'occasion pour moi de me confronter à des questions sur la manière de rédiger une publication (en Anglais qui plus est), aussi bien dans le fond que dans la forme. Le projet autour du taiko a été l'occasion de comprendre les mécanismes de réflexion à développer lorsqu'une problématique est lancée par une personne extérieure au milieu de la recherche. Bien sûr, la collaboration avec un laboratoire d'horizon différente a également été enrichissante.
Culturellement parlant, ce stage aura aussi été une grande découverte. Le Japon étant un pays très riche et très différent du mien, les méthodes de travail, les rapports sociaux, ou simplement la langue ont été pour moi l'objet d'un intérêt constant.
Et si cette expérience de travail a été unique en son genre, elle a davantage confirmé mon envie de poursuivre dans le milieu de la recherche.

Références

[Aucouturier *et al.*, 2010] Aucouturier, J.-J., Johansson, P., & Segnigni, R. 2010. Natural emotional coloring of spoken voices in real-time. *Not yet published.*

[Bertini *et al.*, n.d.] Bertini, G., Fontana, F., Gonzalez, D., Grassi, L., & Magrini, M. Voice transformation algorithms with real time DSP Rapid Prototyping tools. *synthesis*, **6**, 7.

[Bradley & Lang, 1994] Bradley, M.M., & Lang, P.J. 1994. Measuring emotion : The self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, **25**(1), 49–59.

[Bradley & Lang, 1999] Bradley, M.M., & Lang, P.J. 1999. Affective norms for English words (ANEW) : Instruction manual and affective ratings. *University of Florida : The Center for Research in Psychophysiology.*

[Candillier, 2006] Candillier, L. 2006. *Contexte, visualisation et évaluation en apprentissage non supervisé.* Ph.D. thesis, Université Charles de Gaulle, Lille 3.

[Constant *et al.*, 2008] Constant, N., Davis, C., Potts, C., & Schwarz, F. 2008. The pragmatics of expressive content : Evidence from large corpora. *Sprache und Datenverarbeitung.*

[Eide *et al.*, 2004] Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., & Pitrelli, J. 2004. A corpus-based approach to< ahem/> expressive speech synthesis. *In : 5th ISCA Speech Synthesis Workshop.* Citeseer.

[Farner *et al.*, 2008] Farner, S., Veaux, C., Beller, G. Rodet, X., & Ach, A. 2008. Voice transformation and speech synthesis in video games. *In Proceedings of paris game developers conference, paris, france.*

[Pittam Cynthia & Callan, 1990] Pittam Cynthia, J., & Callan, V. 1990. The long-term spectrum and perceived emotion. *Speech Communication*, **9**(3), 177–187.

[Russ *et al.*, 2008] Russ, J.B., Gur, R.C., & Bilker, W.B. 2008. Validation of affective and neutral sentence content for prosodic testing. *Behavior Research Methods*, **40**(4), 935.

[Sabin & Pardo, 2009a] Sabin, A.T., & Pardo, B. 2009a. 2DEQ : an intuitive audio equalizer. *Pages 435–436 of : Proceeding of the seventh ACM conference on Creativity and cognition.* ACM.

[Sabin & Pardo, 2009b] Sabin, A.T., & Pardo, B. 2009b. A method for rapid personalization of audio equalization parameters. *Pages 769–772 of : Proceedings of the seventeen ACM international conference on Multimedia.* ACM.

[Scherer, 2003] Scherer, K.R. 2003. Vocal communication of emotion : A review of research paradigms. *Speech communication*, **40**(1-2), 227–256.

[Scherer & Oshinsky, 1977] Scherer, K.R., & Oshinsky, J.S. 1977. Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, **1**(4), 331–346.

[Slaney, 1993] Slaney, M. 1993. An efficient implementation of the Patterson-Holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep.*

[Slaney, 1998] Slaney, M. 1998. Auditory toolbox, version 2. *Interval Research Corporation*, **10**, 1998.

Annexe

Phrases neutres utilisées pour l'expérience

Im on my way to the meeting
I wonder what that is about
Have you seen him ?
The airplane is almost full
Can you hear me ?
Maybe tomorrow it will be cold
I would like a new alarm clock
Can you call me tomorrow ?
I think I have a doctors appointment
Well stop in a couple of minutes
How did he know that ?
Dont forget a jacket
I think Ive seen this before
The surface is slick

Pages webs développées pour l'expérience

Listening Test: Registration

Welcome to the listening test.

User Agreement

This is the entry point to a listening test. The results will be used for scientific research.

You are asked to log in by giving identification details, which will be kept confidential.
We will use your email address at most twice after the test is done. First, to inform you in the event that you should win a thank-you present in our draw competition. Second, to offer you the opportunity to participate in a follow-up study in a few weeks, which we will ask only once.

If you agree to this, and want to proceed with the test, please fill out the form below.
Note that you need to be at least 18 year-old to take the test.

For any additional information, please contact Jean-Julien Aucouturier (aucouturier-at-gmail.com).

Name: julian

Email address: julian.moreira@julian.com

Submit

Listening Test: User information

Thanks for agreeing to take part in the test. Before we can start, we need to know a bit more about you. This information will be kept confidential, as before.

ABOUT YOU

Please indicate:

- Your age: ____ years old.
- Your gender: ☐ Female ☐ Male
- Whether you are a native english speaker: ☐ Native English ☐ Not Native

---

To proceed to the test, please submit the above information: Submit

Listening Test: Sound Configuration

One last thing before starting: we need you to configurate the volume of audio playback.
Please play the following extract, and set up the volume of playback to a confortable level.
The volume you decide on will be used throughout the test.

Play Stop

Volume control:

When ready, proceed to the test. Next

Listening Test: Tutorial

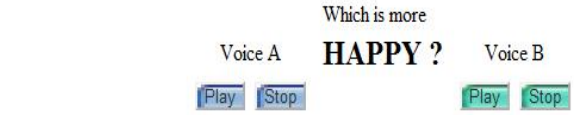
TASK

In this test, you will be asked to listen to English sentences, said in a variety of ways.
For instance, you may hear a sentence said in a "nice" way, or another sentence said in a "sad" way.
We would like you to report on your perception of the "way" the sentences are said.

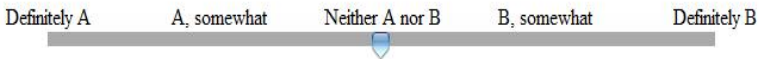
More precisely, you will hear voices by several speakers, saying a variety of English sentences.
For each speaker, you will hear 2 recordings of the same sentence, and you will be asked to determine which of the two sentences best matches a given adjective.
For instance, you may be presented with two voices and asked which one of them best matches the adjective "happy", in your opinion.
In the case where the adjective does not seem to match any of the voices, you will also have the option to say "neither one nor the other voice matches", if you so think.

EXAMPLE

The test will look like the following:



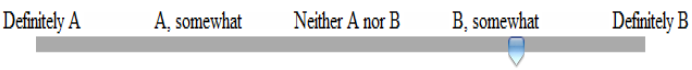
Please rate how much you agree with the above statement on the following scale



EXPLANATIONS:

Use the left side of the scale if you think that the adjective best describes "voice A", i.e. you feel that "voice A" is a lot more like the adjective than "voice B".
Use the right side of the scale if you think that the adjective best described "voice B", i.e. you feel that "voice B" is a lot more like the adjective than "voice A".
Use the middle of the scale if you think that the adjective applies neither to "voice A" nor "voice B", or if you think that it applies equally well to both (i.e. none of them is "more like" the adjective), or if you don't know.

In the case of the example above, if you want to answer e.g. that you think voice B is happier than voice A, you will be able to do so as shown below.



LAST REMARKS:

- For technical reasons, the audio quality of the recordings may not always be perfect. As far as you can, try to ignore the sound quality (noise, etc.) and rate only the way the voice is *spoken* , not *recorded* .
- Please disregard the lexical content of the sentences (i.e. do you like or understand the phrase and what it means). If a sad sentence is spoken in a happy way, please report about the happy *way* , not the sad *words* .
- You may listen to the sounds as many times as needed.

When ready, you can now proceed to the Listening Test.