10ème **C**ongrès **F**rançais d'**A**coustique

Lyon, 12-16 Avril 2010

Le rôle de la prosodie dans la perception de l'effort vocal

Thibaut Fux[1], Gang Feng[2], Véronique Zimpfer[1]

[1]ISL, BP 70043, 68301 Saint-Louis Cedex, France, {thibaut.fux, veronique.zimpfer}@isl.eu
[2]Grenoble Images Parole Signal Automatique, 38402 Saint Martin D'Hères Cedex, gang.feng@gipsa-lab.inpg.fr

Cette étude entre dans le cadre des recherches sur la perception de la distance d'un locuteur uniquement grâce à la perception de l'effort vocal de celui-ci. Si l'effort vocal est aisément perceptible par l'homme, le mécanisme exact de cette perception n'est pas totalement élucidé. L'augmentation de l'effort vocal entraîne des modifications des différents paramètres de la voix, en particulier les paramètres prosodiques et spectraux. L'ensemble de ces paramètres contribue de manière coordonnée à la perception des efforts vocaux mais dans des proportions différentes. Notre étude est destinée à mettre en évidence le rôle prédominant de la prosodie dans la perception de la distance du locuteur par le biais de l'effort vocal. Pour ce faire, un corpus a été enregistré dans une situation de communication de vive voix à distance (5m et 100m), comprenant 3 locuteurs. Le corpus contient 5 phrases ; les enregistrements ont été effectués à 1m des locuteurs. Un vocodeur à prédiction linéaire avec un algorithme de *matching* (DTW) a été implanté permettant ainsi de combiner les paramètres spectraux et prosodiques des deux types de voix. Dans une première voix ainsi synthétisée, on combine la prosodie de la voix de 100m avec les paramètres spectraux de la voix normale (5m) tandis que dans une seconde voix c'est l'inverse. Un test perceptif, basé sur des comparaisons en termes de ressemblance, a été effectué par 18 sujets. Il en ressort que dans 86% des cas, les voix synthétiques basées sur les paramètres prosodiques de la voix à 100m sont identifiées comme les voix distantes (100m), tandis que les voix synthétiques avec les paramètres spectraux de la voix à 100m sont perçues comme étant proche (5m) et ce dans 93% des cas. Ce résultat montre ainsi clairement la prédominance de la prosodie dans la perception de la distance d'un locuteur.

1 Introduction

Cette étude entre dans le cadre des recherches sur la perception de la distance d'un locuteur uniquement grâce à sa voix, plus précisément, grâce à la perception de l'effort vocal de celui-ci.

Bien qu'il soit difficile pour un individu de juger de la distance qui le sépare d'une source sonore quelconque il en est tout autrement lorsqu'il s'agit de la parole. En effet, contrairement à une source sonore quelconque, la voix d'un individu s'adapte et se modifie systématiquement en fonction de la distance qui le sépare de son interlocuteur. On constate naturellement une augmentation de l'effort vocal au fur et à mesure que la distance entre les interlocuteurs croît. On observe généralement que pour une distance proche, le locuteur aura tendance à chuchoter, à parler normalement pour une distance dite conversationnelle et à parler fort, crier ou même hurler pour des distances plus grandes. Ces changements d'efforts vocaux sont aisément perceptibles et associés à des distances différentes [1,2]. La voix constitue ainsi un vecteur riche d'informations pour la perception de la distance étant donné que ces caractéristiques sont systématiquement modifiées avec l'effort vocal. La perception de l'effort vocal est alors directement liée à la perception de la distance. Toutefois les mécanismes exacts de la perception de l'effort vocal ne sont pas totalement élucidés.

L'effort vocal se reflète à travers les variations de plusieurs paramètres. Il existe en effet, un nombre important de paramètres qui se voient altérés par l'augmentation de la force vocale tels que la source vocale (ou source glottique) [3], les formes spectrales des sons (la répartition fréquentielle d'énergie) [4] ou encore la structure et les variations temporelles des énoncés [5].

L'ensemble de ces paramètres peut être schématiquement classé en deux catégories : (1) les *paramètres spectraux*, caractérisant l'enveloppe spectrale d'un son (fréquences des formants, largeurs de bande, déclinaison spectrale), (2) les *paramètres prosodiques* représentant l'ensemble des indications concernant la fréquence fondamentale, l'intensité et la durée locale et globale d'une phrase. Cependant, l'apport de chacune de ces catégories dans la perception de la distance n'a pas encore été clairement défini. Deux études se sont intéressées au rôle de la prosodie dans la perception de la distance et/ou de l'effort vocal.

Brungart et al. [6], ont montré comment les variations de la fréquence fondamentale affectent la perception de la distance d'un locuteur. Sur la base d'enregistrements effectués avec des intensités allant de 60dB SPL à 96dB SPL par pas de 6dB SPL, ils ont « interchangé » le contour de la fréquence fondamentale (notée *f0)* de chacun de ces enregistrements avec celui du niveau d'intensité supérieur (+6dB SPL) ou inférieur (-6dB SPL). A partir d'un test perceptif il en déduit que les voix modifiées en augmentant *la f0 (*en accord avec la voix à +6dB SPL) paraissent provenir de plus loin que la voix de référence et que les voix modifiées en diminuant la *f0* (en accord avec la voix à -6dB SPL) paraissent provenir de plus près. Ainsi les variations de *f0* sembleraient être un facteur important pour la perception de la distance et donc de l'effort vocal.

Tassa et Liénard [7], quant à eux, ont utilisé des techniques de transformation de la voix afin de transformer un niveau d'effort vocal en un autre plus élevé. Malgré l'absence de test perceptif, ils en concluent que les variations d'intensité, de durée phonémique et de fréquence

fondamentale sont des indicateurs de l'effort vocal plus fort que les variations de la fréquence des formants ou encore que la pente spectrale. Toutefois, cette étude a uniquement été réalisée sur la base de voyelles prononcées à des efforts vocaux différents.

Dans cet article nous nous proposons de compléter ces études en jugeant de l'apport de chaque groupe de paramètres (prosodiques et spectraux) pour la perception de l'effort vocal. Dans cet objectif, notre approche consiste à modifier les paramètres d'une voix dite « normale » en lui greffant les paramètres issus d'une voix produite avec un effort vocal plus intense due à la distance. Pour ce faire, nous avons élaboré un corpus en réalisant des enregistrements de voix reflétant l'effort vocal nécessaire à la communication à distance. Après l'analyse des caractéristiques de ces voix, nous montrons comment nous avons transposé les paramètres d'une voix criée sur une voix normale grâce à un vocodeur à prédiction linéaire et un algorithme de *matching* (DTW). Par la suite un test perceptif a permis d'identifier lequel des deux groupes de paramètres joue un rôle prédominant dans la perception de l'effort vocal.

2 Corpus

Le corpus a été enregistré dans une situation de communication de vive voix à distance (5m, 25m, 50m, 75m et 100m). Il contient trois locuteurs masculins. Les enregistrements ont été réalisés en champ libre sur une route en macadam entourée par des champs.

La voix des locuteurs est enregistrée simultanément par deux microphones : l'un situé à 1m du locuteur et l'autre près de son interlocuteur. Cependant, nous n'utilisons dans cette étude que les signaux enregistrés près du locuteur (à 1m) pour les distances de 5m et 100m. En effet, nous considérons qu'avec une faible distance de communication (5 m), la voix du locuteur est une voix normale, alors qu'à 100m, sa voix reflète une augmentation significative des efforts vocaux. Dans la suite de cet article nous désignerons par *voix parlées* les voix produites pour une distance de communication de 5m et par *voix criées* celles produites pour une distance de 100m.

Pour la réalisation de ce corpus, cinq phrases ont été utilisées :

(Ph.1) : Lève le bras. /l ɛ v l ə b r a/
(Ph.2) : Tu me vois? /t y m ə v w a/
(Ph.3) : Tu m'entends? /t y m ɑ̃ t ɑ̃/
(Ph.4) : Bouge la tête. /b u ʒ l a t ɛ t/
(Ph.5) : Viens vers moi. /v i œ̃ v ɛ r m w a/

Afin d'assurer un niveau de communication adéquat à la distance d'enregistrement, un interlocuteur s'est placé à côté du microphone situé à la distance souhaitée. Cet interlocuteur exécutait les actions prononcées par le locuteur (Ph.1, Ph.4 et Ph.5), ou répondait aux questions (Ph.2 et Ph.3) dans le cas où l'effort vocal produit par le locuteur lui permettait de comprendre la phrase.

Les enregistrements ont été effectués par un microphone ½ pouce B&K (type 4189), un conditionneur B&K (type 5935) et un enregistreur numérique portable (M-Audio, Microtrack II) à une fréquence de 16kHz.

3 Analyse des signaux enregistrés

La comparaison entre la voix prononcée pour une distance de communication de 100m et la voix parlée (5m) nous permet d'étudier les variations de différents paramètres reflétant l'effort vocal, tel que (1) l'intensité, (2) la fréquence fondamentale *f0*, (3) les durées phonémiques, (4) les fréquences et largeurs de bandes des formants. L'ensemble des valeurs présentées dans ce paragraphe ont été obtenues à l'aide du logiciel *Praat* [8]. Les différents enregistrements sont nommés suivant le schéma LxPHy : Lx représente le locuteur et PHy la phrase prononcée. x et y représente respectivement le numéro du locuteur et de la phrase.

3.1 Intensité (*I*)

La Figure 1 donne les valeurs de l'intensité moyenne et de l'écart type pour chaque locuteur et chaque phrase pour les deux situations de communication (5m et 100m).

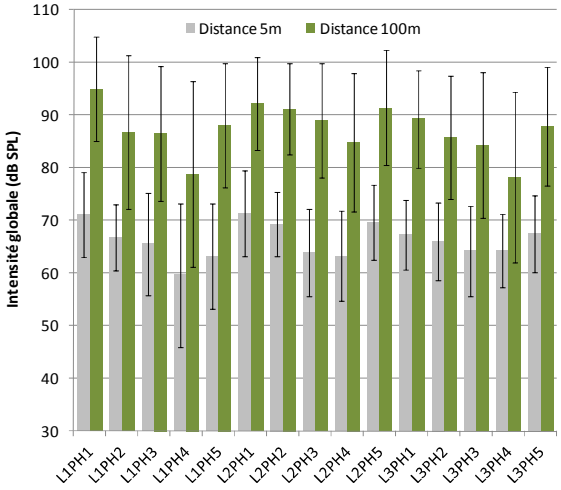

Figure 1 : Intensités moyennes des enregistrements dans les deux situations de communication. Les barres verticales représentent les écarts types.

Il existe une différence significative entre les valeurs moyennes de *I* ($p<0,001$) pour les voix parlées et les voix criées. En moyenne les voix parlées ont une intensité de 62,1dB SPL tandis que les voix criées ont une intensité de 82dB SPL. On observe une augmentation moyenne de l'intensité de 19,9dB entre la voix normale et la voix permettant de communiquer à une distance de 100m, ce qui correspond à une augmentation de +4,6dB par doublement de la distance.

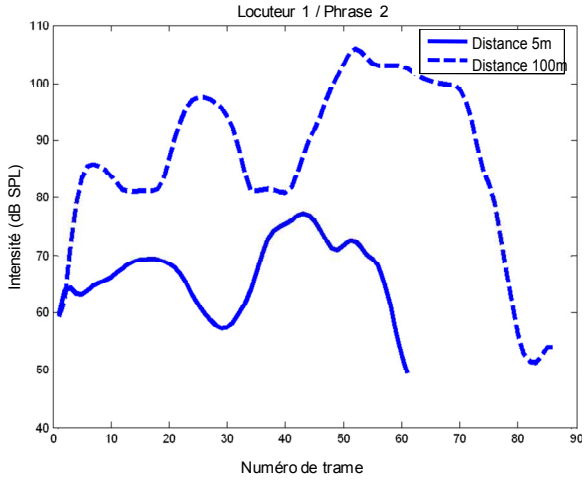

Figure 2 : Variation de l'intensité pour la phrase 2 du locuteur 1, pour une distance de 5m et 100m.

Toutefois, le passage de la courbe d'intensité pour une voix parlée à la courbe d'intensité pour une voix criée n'est pas simplement régi par une loi proportionnelle. En effet, on peut observer sur la Figure 2 des différences quant à l'allure générale de la courbe d'intensité. En effet, comme Rostolland [9], on observe une augmentation d'intensité moins élevée pour les consonnes que pour les voyelles. Ceci étant on constate une dynamique d'intensité plus grande pour les voix criées.

3.2 Fréquence fondamentale (*f0*)

La Figure 3 donne les valeurs de la fréquence fondamentale moyenne et l'écart type pour chaque locuteur et chaque phrase pour les deux situations de communication (5m et 100m). La Figure 4 représente les étendues $\Delta f0$ qui sont définies d'après l'équation (1).

$$\Delta f0 = f0_{max} - f0_{min},\ pour\ f0_{min} > 0 \qquad (1)$$

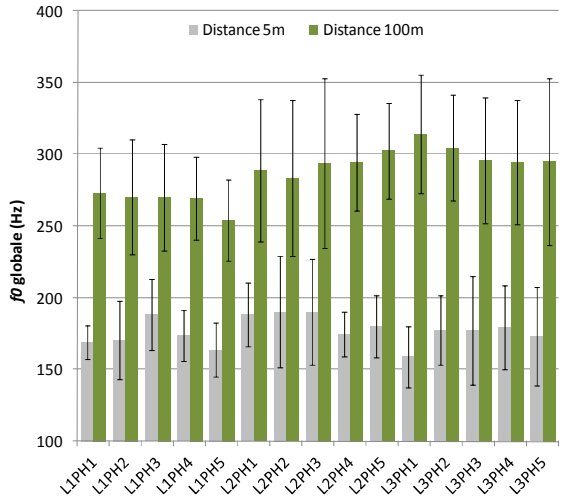

Figure 3 : Fréquences fondamentales moyennes des enregistrements dans les deux situations de communication. Les barres verticales représentent les écarts types.

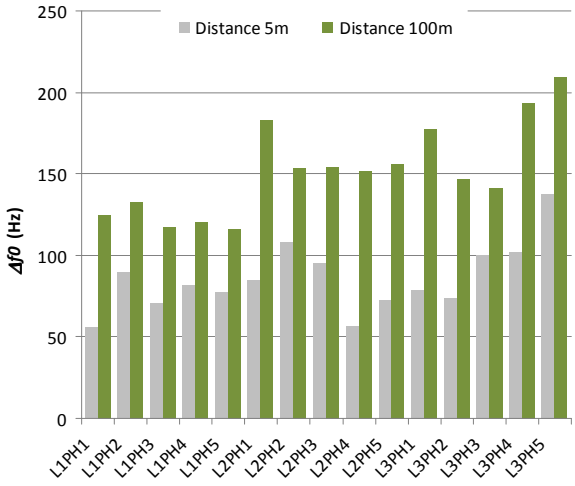

Figure 4 : Etendues de la *f0* des enregistrements dans les deux situations de communication.

Il existe une différence significative entre les valeurs moyennes de *f0* ($p<0,001$) et les dynamiques ($p<0,001$) entre les voix parlées et les voix criées. En moyenne les voix parlées ont une *f0* moyenne de 177Hz et une étendue de 85.6Hz tandis que les voix criées ont une *f0* moyenne de 287Hz et une étendue de 152Hz.

De la même façon que pour les courbes d'intensité, le passage de la courbe de *f0* pour une voix parlée à la courbe de *f0* pour une voix criée n'est pas simplement régi par une loi proportionnelle. En effet, on peut observer sur la Figure 5 des différences quant à l'allure générale du contour de *f0*.

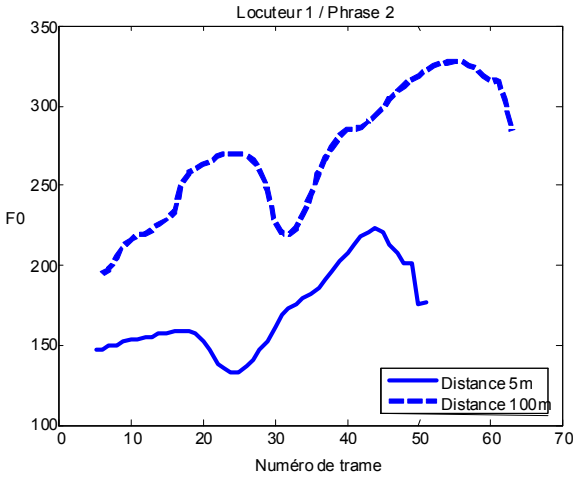

Figure 5 : Variation des contours de *f0* pour la phrase 2 du locuteur 1, pour une distance de 5 m et 100 m.

3.3 Durées

L'ensemble des phrases ont une durée plus élevée quand l'effort vocal est grand (voix à 100m). En moyenne on observe une augmentation de 20% de la durée des phrases, excepté pour la phrase 5 du locuteur 1 pour laquelle la durée diminue de 4%. Cette augmentation est observée sur la Figure 2 et 5.

Voyelles			*Consonnes*			
	Δt	***σt***	**Famille**		***Δt***	***σt***
a	20,71%	8,15%	*Liquide*	l	-14,06%	17,68%
ə	25,10%	28,11%		ʀ	6,96%	35,48%
i	-20,02%	8,08%	*Fricative sonore*	v	10,35%	43,83%
y	26,45%	21,79%		ʒ	10,20%	0,69%
u	52,65%	48,57%	*Occlusive sonore*	b	-22,65%	25,17%
ɛ	51,70%	31,93%	*Occlusive sourde*	t	-13,75%	49,30%
ɑ̃	34,42%	12,65%	*Nasale*	m	23,92%	43,24%
ɛ̃	13,55%	32,30%	*Semi consonne*	w	30,75%	46,27%

Table 1 : Variation de la durée des phonèmes entre la voix criée (100m) et la voix parlée (5m).

Les variations de durée des différents phonèmes contenus dans les phrases utilisées sont présentées dans la Table 1. On observe une augmentation de 32% en moyenne de la durée des voyelles entre une voix parlée et une voix criée (hormis pour le phonème /i/). On observe pour les voyelles finales une augmentation de 54% pour le phonème /a/ des phrases Ph.1, Ph.2 et Ph.5 et une augmentation de 58% pour le phonème /ɑ̃/ de la phrase Ph.3. Ces résultats sont en accord avec Rostolland [7] qui montre que la dernière voyelle d'une phrase est plus allongée que les autres pour des voix criées (66% pour la voyelle finale comparé à 33% pour des « non-finales »). A l'inverse, pour la consonne finale /t/ de la phrase Ph.4, on trouve une

diminution de la durée de -48%. De la même manière on observe une augmentation de la durée des consonnes « non-occlusives » (hormis pour le phonème /l/) de 16%. *A contrario*, la durée des occlusives tend à diminuer (-18%). On remarque toutefois une grande variabilité de ces valeurs. Ceci peut provenir de la différence de stratégie de communication utilisée par les locuteurs. En effet, Garnier [10] a montré qu'il existe des stratégies d'adaptation à l'effort vocal différentes en fonction des sujets. Cette variabilité peut également provenir du faible nombre de locuteur ou encore du corpus qui n'est pas adapté à ce type de mesure.

3.4 Fréquences et largeurs des Formants (f_x, b_x)

Ces relevés ont été effectués à partir de voyelles isolées (a, ə, i, o, y, u) prononcées au cours de l'enregistrement du corpus. La Table 2 donne les valeurs moyennes des déplacements des fréquences formantiques ainsi que des largeurs de bandes pour chaque voyelle.

	Δf_1	Δf_2	Δb_1	Δb_2
a	10,09%	14,75%	-11,26%	-27,60%
ə	27,51%	-4,69%	-13,35%	187,39%
i	-6,33%	-2,75%	20,23%	34,96%
o	36,75%	5,08%	-87,90%	-17,51%
y	11,70%	-2,01%	-29,64%	13,10%
u	2,29%	0,30%	85,25%	-75,27%

Table 2 : Evolution des fréquences et largeur de formants entre les voix parlées et les voix criées

Hormis pour le phonème /i/, on constate une augmentation de la fréquence du premier formant (f_1) de 17.7%. De manière générale (hormis pour le phonème /a/) la fréquence du deuxième formant (f_2) ne varie pas ou très peu. Ces observations vont dans le sens de Liénard [11]. Cependant les largeurs de bande de f_1 (b_1) semblent diminuer tandis que pour les largeurs de f_2 (b_2) on n'observe pas de tendance particulière.

4 Synthèse et modification de la voix

Le paragraphe précédent nous a permis de mettre en évidence les ampleurs des modifications résultant de l'augmentation de l'effort vocal. Toutefois, ces modifications prises séparément ne peuvent refléter complètement leurs impacts dans la perception des efforts vocaux. C'est pourquoi il nous a paru nécessaire d'effectuer des tests perceptifs avec des voix synthétiques obtenues à partir d'une voix parlée mais dans lesquelles une partie des paramètres (prosodiques ou spectraux) provient d'une voix criée. Pour ce faire, nous avons développé un vocodeur à prédiction linéaire ainsi qu'un algorithme de *matching*.

4.1 Vocodeur à prédiction linéaire

Le vocodeur utilisé dans cette étude est un vocodeur basé sur la prédiction linéaire [12]. L'enveloppe spectrale du signal de la parole est modélisée à l'aide d'un modèle autorégressif (AR) caractérisé par les coefficients de prédiction a_i. Pour chaque phrase prononcée, on extrait ces coefficients toutes les 10ms (taille de la trame). Pour chaque trame, on estime aussi la fréquence fondamentale (si la trame est voisée) et son intensité.

La restitution du signal est réalisée à l'aide du même modèle (cf. Figure 6) : un filtre AR d'ordre p est utilisé pour reproduire l'enveloppe spectrale, qui est alors excité soit par un train d'impulsions (pour la parole voisée) soit par un bruit blanc (pour la parole non-voisée).

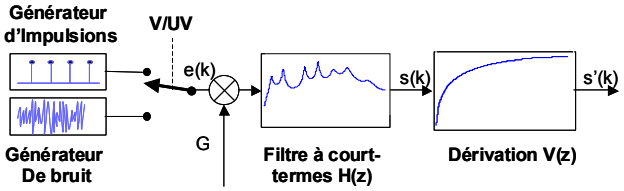

Figure 6 : Schéma bloc du vocodeur à prédiction linéaire

La reconstruction du signal est effectuée suivant la technique de l'*overlapping-add* (addition recouvrement) afin de réduire les discontinuités engendrées par ce genre de synthèse par trame. Concrètement cela signifie que la fenêtre d'analyse est plus longue que son pas d'avancement. La fenêtre d'analyse utilisée a une longueur de 20ms pondérée par une fonction de *Hanning* et est décalée par pas de 10ms (soit un recouvrement de 50%). Le prédicteur H(z) est un filtre AR tout-pôle d'ordre p=18. Le coefficient a du dérivateur ($V(z)=1-az^{-1}$) représentant la radiation aux lèvres est égal à 0,98.

Pour améliorer la qualité de la synthèse, les différents paramètres utilisés sont lissés entre deux trames successives, en particulier pour la période de la fondamentale. Ce lissage consiste à faire varier progressivement la période fondamentale d'une trame à l'autre selon une loi linéaire.

Malgré sa simplicité, ce vocodeur permet de manipuler aisément et séparément les paramètres spectraux et prosodiques de la parole. Certes, ce vocodeur ne permet pas d'obtenir une très grande qualité de synthèse mais nous estimons que la qualité de la parole synthétisée est suffisante pour notre étude.

4.2 DTW (Dynamic Time Warping)

Dans notre étude, nous cherchons à modifier l'une des catégories de paramètres (prosodique ou spectrale) d'une voix normale en utilisant ceux d'une voix criée. Par exemple, s'il s'agit de paramètres spectraux, il consiste à prélever ces paramètres dans la voix criée et de les « greffer » dans la voix normale.

Compte tenu des structures temporelles différentes des deux voix, il est indispensable de réaliser un alignement temporel. Pour ce faire, nous utilisons la technique appelée *Dynamic Time Warping* (DTW) ou déformation temporelle dynamique. Dans cette méthode, la similarité entre les différentes trames des deux voix est assurée à l'aide du calcul d'une distance spectrale. Le meilleur alignement consiste à trouver une correspondance entre ces trames minimisant les distances spectrales.

Concrètement cette méthode consiste à segmenter les deux signaux ($a(n)$ et $b(n)$) en fenêtres d'analyses et de calculer une distance spectrale locale d pour l'ensemble des combinaisons possibles entre ces fenêtres (cf. équation (2)). Une distance spectrale minimale signifie un maximum de ressemblance spectrale.

$$d(i,j)=\frac{\sum_{k=1}^{N/2}\left|A_i(k)\right|.\left|B_j(k)\right|}{EA_i.EB_j} \quad \text{i=1,…,I ; j=1,…,J.} \quad (2)$$

Dans l'équation (2), I et J sont respectivement le nombre total de fenêtre d'analyses des signaux *a(n)* et *b(n)*. A_i et B_j sont respectivement les transformées de Fourier (FFT) des i^{ème} et j^{ème} fenêtres d'analyses de *a(n)* et *b(n)*. *EB* et *EA* sont les énergies de A_i et B_j et N la longueur de la FFT.

L'équation (3) calcule ainsi une matrice de distance spectrale croisée. La distance globale optimale D, modélisant le meilleur alignement, est obtenue en cherchant le chemin dans cette matrice qui minimise la somme des distances locales pour aller d'un point initial ($i=I, j=J$) à un point final ($i=1, j=1$). On peut facilement démontrer que la somme des distances minimales locales est égale à la distance minimale globale. Ainsi le chemin optimal D se calcule de la manière suivante :

$$D(i,j) = D(i,j) + \min[d(n-1,j), d(n,j-1), d(n-1,j-1)] \quad (3)$$

De cette manière, on contraint le chemin D à suivre un trajet « monotone » et « plausible » tout en suivant le chemin le plus court.

Sur la Figure 7, on peut voir les deux signaux à aligner (*a(n) :* voix parlée, et *b(n) :* voix criée), la matrice des distances spectrales croisées *d,* et représenté par les croix blanches le chemin optimal D.

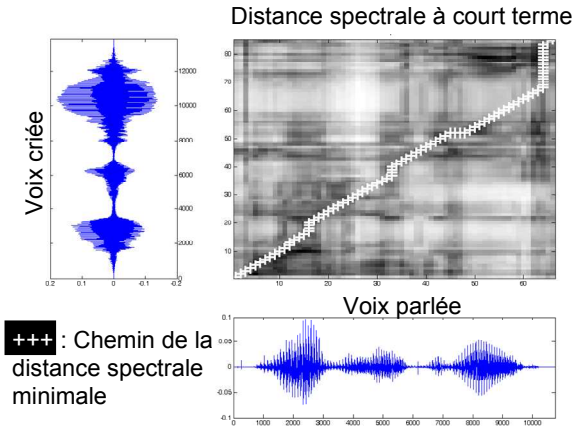

Figure 7 : Exemple de DTW

Ainsi, grâce au vecteur D il est possible d'associer une fenêtre d'analyses de *a(n)* à son équivalent spectral (le même son) sur le signal *b(n)*.

4.3 Génération des fichiers tests

Les fichiers du test sont générés de la façon suivante : pour tester le rôle des paramètres spectraux d'une voix criée sur une voix parlée, on conserve la prosodie de celle-ci (structure temporelle, évolution de la *f0*, évolution de l'intensité) et on lui greffe les paramètres spectraux de la voix criée en utilisant les correspondances établies avec la DTW. Après la synthèse on obtient une voix mixte appelée *Smod_spec* (cf. Figure 8). Pour tester le rôle des paramètres prosodiques d'une voix criée, on utilise ces paramètres pour la synthèse tout en combinant les paramètres spectraux initiaux à la voix normale (toujours via la DTW). La synthèse nous donne une voix mixte appelée *Smod_pros*.

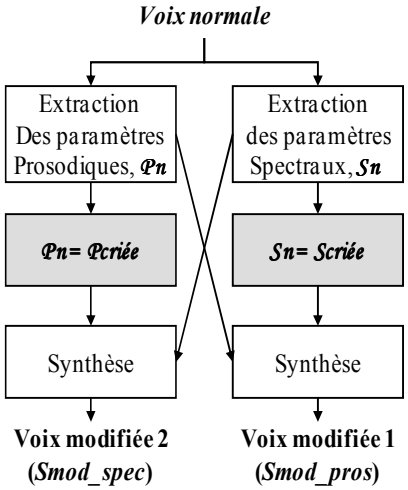

Figure 8 : Principe de la modification de la voix par « *matching* », *Sn* et *Scriée* sont respectivement les paramètres spectraux de la voix parlée et de la voix criée. *Pn* et *Pcriée* correspondent aux paramètres prosodiques de la voix parlée et de la voix criée.

5 Test perceptif

Un test perceptif a été organisé pour étudier l'apport de chaque jeu de paramètres dans la perception de l'effort vocal. Il s'agit de comparaison en termes de ressemblance. Les voix d'origines sont également re-synthétisées afin d'être insérées dans le test. L'ensemble des signaux ont été égalisés en intensité pour ce test.

5.1 Protocole de test

Pour un locuteur donné et une phrase donnée deux voix modifiées ont été créées (*Smod_pros, Smod_spec*) à partir des enregistrements de la voix parlée et de la voix criée du locuteur pour la même phrase. Le test consiste alors à comparer chaque voix modifiée aux voix d'origines à partir desquelles elle a été créée.

Le test se présente de la manière suivante : on fait entendre aux sujets successivement, et une unique fois, la voix parlée, la voix modifiée 1 puis la voix criée. Un deuxième triplet de stimuli est également entendu pour la voix modifiée 2. Ainsi, nous obtenons deux triplets de stimuli pour chaque locuteur et pour chaque phrase. Ceci constitue une base de trente triplets à comparer. Lors de l'écoute, les phrases ont été séparées par une pause de 300ms. L'ensemble des trente triplets de stimuli ont été comparés par chaque sujet dans un ordre d'apparition aléatoire. Les sujets devaient alors, pour chaque triplet de stimuli, répondre à la question : à laquelle des deux références ressemble le plus la phrase modifiée?

Ce test perceptif s'est déroulé dans une chambre d'audiométrie afin que les sujets ne puissent être perturbés par du bruit ambiant. Il a été réalisé par dix-huit sujets (15♂, 3♀) dont la répartition des âges est la suivante : six sont âgés de moins de 30ans, cinq sont âgés de 30 à 40ans, six de 40 à 50ans et un est âgé de plus de 50ans.

5.2 Résultats

La Table 3 donne les résultats du test subjectif des dix-huit sujets. La première voix modifiée (*Smod_pros*) est clairement perçue comme étant très similaire à une voix criée sans doute du fait qu'un effort vocal significatif est

perçu. Cette voix est alors perçue comme une voix produite dans le but de pallier une grande distance (*Voix criée*) dans 86% des cas. La seconde voix synthétisée (*Smod_spec*) ne donne pas la même perception de l'effort vocal. Elle est plutôt perçue comme une voix produite dans un but de communication à faible distance (*Voix parlée*) dans 93% des cas.

	Voix parlée	**Voix criée**
Smod_pros	14%	86%
Smod_spec	93%	7%

Table 3 : Ressemblance des phrases modifiées par rapport aux phrases de référence.

Le facteur *Locuteur* ne semble pas avoir un rôle déterminant pour ces résultats. Bien que les valeurs soient différentes pour les trois locuteurs, les ordres de grandeurs ainsi que la tendance globale des résultats sont similaires.

6 Discussion

De toute évidence, les voix modifiées manquent de naturel. En effet, la perte de qualité due au synthétiseur, combinée aux modifications (prosodiques ou spectrales), engendrent des voix qui ne semblent pas être produites par l'être humain. Les paramètres prosodiques et spectraux étant étroitement liés, modifier un seul jeu de paramètres génère des phrases qui ne sont pas naturelles. Les voix obtenues restent néanmoins parfaitement intelligibles et ne sont pas désagréables à entendre. De plus, les sujets qui ont participé à l'expérience ont jugé ne pas avoir été perturbés par l'écoute des voix modifiées. De ce fait, nous pensons que la qualité des voix synthétisées n'a pas biaisé les résultats obtenus.

Ces résultats montrent que les paramètres spectraux à eux seuls, contrairement aux paramètres prosodiques, ne contribuent pas de manière significative à la perception des efforts vocaux. Ainsi, nous pouvons dire que les paramètres prosodiques sont prédominants dans la perception des efforts vocaux. Ceci ne nie pas pour autant la contribution globale des paramètres spectraux. En effet, leur apport n'est pas à négliger du fait que l'effort vocal ne se reflète qu'à travers un tout.

Rappelons que notre objectif n'est pas de quantifier de manière absolue l'apport de chacun de ces groupes de paramètres, mais que nous cherchons à mettre en évidence l'apport relatif à chacun d'eux. Ces résultats nous indiquent clairement qu'une plus grande prise en considération des variations prosodiques par rapport aux variations spectrales est nécessaire pour la perception de la distance par le biais de l'effort vocal.

7 Conclusion

Nous montrons le rôle prédominant des paramètres prosodiques dans la perception de l'effort vocal en utilisant un test perceptif à partir de voix parlées modifiées. Les différents paramètres (prosodiques et spectraux) d'une voix criée, correspondant à une situation de communication à distance, ont été greffés sur une voix parlée afin de tester leur pertinence dans la perception de l'effort vocal. La modification de la voix parlée est effectuée à l'aide d'un vocodeur à prédiction linéaire et un algorithme de *matching* (DTW). Il en ressort que dans 86% des cas, les voix synthétisées basées sur les paramètres prosodiques de la voix à 100m sont identifiées comme étant analogues aux voix distantes d'origine, tandis que les voix synthétisées à partir des paramètres spectraux de la voix à 100m sont perçues comme étant proches des voix parlées et ce dans 93% des cas. Ce résultat montre ainsi clairement la prédominance de la prosodie dans la perception de la distance d'un locuteur.

Remerciements

Les auteurs tiennent à remercier tout particulièrement l'ensemble des personnes ayant participé aux enregistrements et au test de perception, ainsi que M. Karl Buck et M. Sébastien De Mezzo pour leur aide et leurs conseils précieux.

Références

[1] Fux Th., Zimpfer V., "Spatialisation du son : Corrélation entre la distance perçue et l'effort vocal", *ISL-RV 218/2009, Institut franco-allemand de Recherches de Saint-Louis (ISL), France,* (2009).

[2] Brungart, D.S. and Scott, K.R., "The effects of production and presentation level on the auditory distance perception of speech", *J. Acoust. Soc. Am.*, 110, 425-440 (2001).

[3] Holmberg E.B., Hillman R.E., Perkell J.S., "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice", *J. Acoust. Soc. Am*., 84(2), 511-529, (1988).

[4] Rostolland D., "Acoustic features of shouted voice", *Acustica*, 50, 118-125, (1982).

[5] Traunmüller H., Eriksson A., "Acoustic effects of variation in vocal effort by men, women, and children", *J. Acoust. Soc. Am*., 107, 3438-3451, (2000).

[6] Brungart D.S., Kordik A.J., Das K. and Shaw A.K., "The effects of f0 manipulation on the perceived distance of speech", *Proc. of the international conference on spoken language processing, Denver, CO,* 1641-1644, (2002).

[7] Tassa A., Liénard J.S., "A new approach to the evaluation of vocal effort by the PSOLA method", *The European Student Journal of Language and Speech, http://www.essex.ac.uk/web-sls/papers/00-01/00-01.html*, (valide en février 2010), (2000).

[8] Boersma P., "Praat, a system for doing phonetics by computer", *Glot. International*, 5, 341-345, (2001).

[9] Rostolland D., "Phonetic features of shouted voice", *Acustica*, 51, 80-89, (1982).

[10] Garnier M., "Communiquer en environnement bruyant : de l'adaptation jusqu'au forçage vocal." *Thèse de doctorat, Université de Paris 6,* (2007).

[11] Liénard J.S., Di Benedetto M.G., "Effect of vocal effort on spectral properties of vowels", *J. Acoust. Soc. Am*. 106, 411-422, (1999).

[12] Makhoul, J. "Linear prediction: a tutorial review", *Proc. IEEE*, 63, 561-580, (1975).