Timbre local, timbre global et cohérence du timbre : l'éclairage de la perception de la voix

Blas Payri

Timbre local, timbre global et cohérence du timbre : l'éclairage de la perception de la voix

Blas Payri

Groupe Traitement du Langage Parlé

LIMSI –CNRS, BP 133 91403 Orsay, France

Tél.: ++33 (0)169 85 80 67 - Fax: ++33 (0)169 85 80 88

Mail: blas@limsi.fr - http://www.limsi.fr/Individu/blas

Mots clés : perception du timbre, timbre vocal, qualité de voix, psychoacoustique

RÉSUMÉ

Cet article présente des expériences sur le timbre vocal, pour dégager des notions et des méthodes expérimentales utiles en perception du timbre en général. Nous dégageons la notion de timbre global et timbre local, en concluant que pour une phrase (musicale) il y a une perception globale du timbre, mais que chaque note peut avoir des perceptions locales très différentes. Nous définissons ensuite la notion de cohérence du timbre, qui étudie les variations admissibles de timbre local pour que les différentes notes (ou syllabes) semblent provenir de la même source.

1 INTRODUCTION

Nous exposons dans cet article des résultats concernant la perception de la voix parlée, dans le but de dégager des cadres expérimentaux pour la perception du timbre en général.

1.1 Le timbre pour la voix et les instruments

Tout d'abord, il convient de se demander si la notion de timbre s'applique aux mêmes mécanismes perceptifs quand nous traitons la voix ou les instruments, d'autant plus que les définitions du timbre demeurent instables. Pour Schaeffer [Sch66], le timbre est « ce à quoi on reconnaît que divers sons proviennent d'un même instrument », mais Schaeffer parle aussi « du timbre d'un son en le considérant comme une caractéristique propre de ce son, perçue pour elle-même ». Castellengo [Cas94] étend la notion de timbre pour la voix : *« Qu'est-ce que le* « timbre » *de la voix ? La voix est un instrument permettant de produire à volonté des variations de hauteur et des variations spectrales. [...] Ceci étant posé, il faut remarquer que le mot* timbre *peut revêtir des sens très différents. En effet nous parlons du timbre de la voix d'une personne en le comparant à celui d'une autre personne, ou encore des différents timbres d'une même personne, ou encore du timbre de la voix parlée par comparaison à celui de la voix chantée. Il est question aussi du timbre de chaque voyelle, quel qu'en soit le locuteur. Une analyse des situations montre que l'on peut faire correspondre à ce terme deux sortes de caractéristiques acoustiques : 1) celles qui permettent d'identifier la personne (timbre de la voix de Mme X, qu'elle soit enrhumée ou en pleine santé), 2) celles qui nous permettent de qualifier les sons, principalement dans le domaine spectral, lorsqu'on parle de voix claire, sourde, perçante, nasale, etc. Dans ce cas nous préférons employer le terme de « couleur » ou de « sonorité » de la voix, réservant le mot timbre pour la première situation. »*

1.2 Les recherches sur le timbre

Les expériences discutent rarement si le timbre qu'elles cherchent à étudier est une description abstraite (sonorité) ou la caractérisation d'une source. Dans le domaine du timbre instrumental, la recherche a été marquée par les travaux de Grey [Gre75], puis de Wessel [Wes79]. Ces recherches se basent sur la définition de l'Acoustic Society of America : « Le timbre est l'attribut de la sensation auditive suivant lequel un auditeur peut différencier deux sons présentés dans les mêmes conditions et ayant la même sonie et la même

hauteur ». Nous voyons que cette définition se place du côté de la sonorité propre d'un son (au sens de Castellengo), indépendamment de la source. Dans ces recherches, on utilise comme matériau des sons tenus de différents instruments, avec même sonie, même hauteur et même durée. Les auditeurs doivent indiquer la similarité globale pour chaque paire de sons (écoute holistique), ce qui permet de définir une matrice de distance qui est ensuite analysée pour créer un espace multidimensionnel de timbres. Notons que ces recherches n'utilisent qu'un échantillon par instrument, ce qui implique que le timbre est ce qui différencie globalement les instruments, et que la sonorité d'un seul échantillon suffit à caractériser l'instrument duquel il provient.

L'écoute holistique a été employée pour la perception des voix pathologiques (Kreiman et Gerratt [Kre96]), et des voix saines (Walden [Wal78], et les travaux pionniers de Murry et alii [Mur78, 80]) Le but de ces expériences est de disposer d'une description de l'espace perceptif sans a priori avec l'obtention de paramètres et de stratégies perceptives " proches de l'écoute quotidienne de tout un chacun " ([Wal78]). Ces expériences posent que le timbre est la distance holistique (ou de similarité globale) entre sons. Cependant on note une confusion entre sonorité abstraite des sons et caractéristiques sonores d'une source.

Une deuxième classe d'expériences cherchant à définir les caractéristiques sonores utilisent l'écoute par axes : on part d'axes perceptifs prédéfinis, dont les auteurs cherchent à mieux comprendre les corrélats acoustiques, la validité, et les interdépendances entre axes.

Dans le domaine de l'identification du locuteur, on peut se référer aux travaux de Voiers [Voi64], qui avait travaillé à partir d'une centaine de descripteurs (paires d'adjectifs opposés) pouvant décrire les voix des locuteurs : les auditeurs devaient juger les paires de voix selon chaque critère, et après une analyse en facteurs principaux, on ne sélectionnait que l'ensemble minimal suffisant de critères permettant de distinguer tous les locuteurs (ces expériences n'ont pas été satisfaisantes car les critères perceptifs sont très instables, et le nombre d'échantillons insuffisant). La recherche en pathologie vocale utilise énormément cette approche : le but est de créer un outil perceptif d'évaluation des pathologies vocales qui soit objectif et complet, valable pour différents groupes d'auditeurs experts (voir les travaux de Kreiman [Kre96]).

1.3 Les variations de qualité vocale à l'intérieur d'une phrase

On peut diviser les différences de timbre entre différents échantillons comme provenant d'une diversité inter-sources (les caractéristiques physiques des instruments et des locuteurs qui contraignent les sons que ceux-ci peuvent émettre), ou d'une diversité intra-source (les variations de dynamique, d'attaque, de bruit… qui peuvent être produites par un locuteur ou un instrument). L'essentiel des expériences décrites dans la littérature vocale et instrumentale n'étudient que les variations inter-source : chaque échantillon représente un et un seul locuteur, et la perception des différences entre échantillons de voix est considérée comme équivalente à la perception des différences entre locuteurs.

Dans les variations intra-source, on peut distinguer les différences globales entre deux phrases (par exemple, une phrase jouée *piano* par rapport à une autre jouée *forte*), et les différences à l'intérieur même d'une phrase. En effet, la prosodie est une variation des contours mélodiques mais aussi une variation de timbre au sein de la phrase. Sluijter et van Heuven [Slu96] montrent que la pente spectrale est un indice très important dans la perception de l'accent tonique en anglais et en néerlandais. Les recherches de Klatt [Kla90] montrent que tout au long d'une phrase, les manifestations acoustiques du souffle varient considérablement et tendent à augmenter pour des syllabes atones. Beaucoup d'élocutions tendent à finir sur une type de vibration « soufflé-laryngé ». Finalement, d'autres irrégularités de voisement apparaissent fréquemment, spécialement en fin de phrase.

Dans les expériences que nous décrivons dans cet article, nous étudions les variations des caractéristiques sonores à l'intérieur même d'une phrase. Nous cherchons d'abord à comprendre si les différents échantillons provenant d'une même source ont les mêmes caractéristiques perçues (expérience de classification libre et écoute holistique). Ensuite, nous essayons de

comprendre les caractéristiques qui varient entre la perception globale de la phrase et les variations à l'intérieur de la phrase (expérience d'écoute par axes). Finalement, nous essayons d'établir les limites des variations admises dans une phrase, à l'aide de montages de syllabes de différents locuteurs.

2 UN MATÉRIAU SONORE POUR COMPARER L'ÉCOUTE GLOBALE ET LOCALE

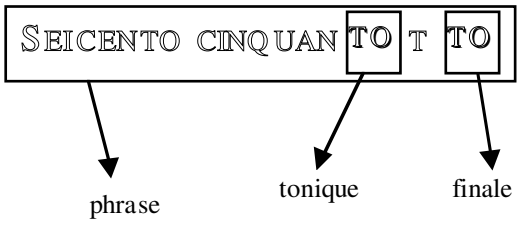

Figure 1 Le matériau sonore était composé de la phrase italienne "seicento cinquantotto" (signifiant "six cents cinquante-huit"), dont on extrait les deux dernières syllabes "to".

Nous avons choisi d'utiliser la base EUROM pour l'italien. Les enregistrements de cette base ont pour avantage d'avoir était faits dans des conditions identiques pour tous les locuteurs, suivant un protocole strict. Par ailleurs tous les locuteurs adoptaient un niveau neutre de force de voix, avec un ton de lecture neutre pour tous les locuteurs, et un voisement normal. Ce type de bases permet de se focaliser sur les différences entre locuteurs.

Pour chaque phrase, nous avons extrait deux syllabes (une tonique, l'autre finale), comme illustré dans la **Figure 1**. Nous disposons ainsi, pour chaque locuteur, d'un matériau long (phrase) qui permettra une perception globale, et deux échantillons courts (syllabes) qui obligeront l'auditeur à une perception locale. Ce choix se justifie ainsi :

> La syllabe est une unité insécable de parole : on peut artificiellement segmenter une syllabe en phonèmes ou diphones, mais un locuteur humain produira toujours une syllabe au minimum (la syllabe la plus simple étant composée d'une voyelle). Certaines expériences utilisent un matériau de type syllabe : par exemple [Wal78] utilisent le mot monosyllabique " beans " dans une expérience de reconnaissance du locuteur. Si on se réfère à la notion d'objet sonore développée par Schaeffer (qui est l'élément minimum pour étudier le son), on constate que la syllabe a les caractéristiques des objets bien formés musicaux, disposant d'une attaque (consonne, ou début de la mise en vibration des cordes vocales pour les syllabes commençant par une voyelle), une transition, avec en particulier des mouvements de formants et de hauteur, un corps (la voyelle) et une chute (soit une occlusion due à une consonne finale, soit la " mort naturelle " de la voyelle).
>
> On peut considérer que la phrase que nous avons prise est d'une longueur suffisante pour que l'auditeur puisse se faire une image globale du locuteur. Par exemple, Schmidt-Nielsen et Stern [Sch85] montrent que la reconnaissance du locuteur s'améliore asymptotiquement avec la durée de l'échantillon, pour atteindre un maximum vers 2 ou 3 secondes, ce qui est la durée d'une phrase courte.

20 locuteurs ont été choisis dans la base : 10 femmes notées F1 à F10, et 10 hommes notés H11 à H20. Parmi les 20 phrases " seicento cinquantotto ", 30 syllabes ont été. Une expérience préliminaire a montré que l'ensemble des 20 syllabes toniques et 20 syllabes finales était trop volumineux pour pouvoir être traitable dans une expérience de classification libre, en effet les auditeurs doivent tenir compte du matériau dans son ensemble pour décider des classes à faire. Nous avons choisi d'éliminer les syllabes qui présentaient le plus de variation, notamment les syllabes finales dévoisées, pour retenir 18 des 20 syllabes toniques et 12 des 20 syllabes finales.

Le matériau que nous utilisons rencontre des restrictions : il est utile pour notre expérience d'étude préliminaire des niveaux de perception du timbre, mais par contre, il est insuffisant si nous visons une étude exhaustive du timbre de la voix. Nous disposons en effet de peu d'éléments : un ensemble de 20 locuteurs différents est forcément peu représentatif de l 'espace du timbre. Par ailleurs, il y a peu de diversité intra-locuteur : il s'agit d'une lecture neutre de chiffres, ce qui restreindra le nombre de dimensions. Il faut aussi noter que nous avons des auditeurs francophones sur un matériau italien qui peuvent percevoir les qualités subjectives de façon différente des

auditeurs italiens. Cependant certaines recherches montrent que des facteurs perceptifs peuvent ne pas dépendre de la langue du locuteur et de l'auditeur : par exemple Braun et Cerrato [Bra99] ont comparé les réponses d'auditeurs allemands et italiens sur des phrases en allemand et en italien, pour montrer qu'il n'y avait pas de différences significatives dans les estimations d'âge ni pour les groupes d'auditeurs ni pour les groupes de locuteurs.

Les limitations de taille de notre matériau font que nous n'obtiendrons pas une étude complète de l'espace perceptif de chaque caractéristique (il faudrait des centaines d'échantillons pour étudier seulement l'âge, le genre, le souffle…). Nous ne cherchons donc qu'à mettre en lumière des phénomènes de variation de perception pour les syllabes et les phrases du même locuteur, qu'il faudrait ensuite approfondir.

3 EXPÉRIENCE D'ÉCOUTE HOLISTIQUE

Etant donné que le timbre est étudié à partir de la similarité globale entre échantillons, nous avons commencé notre recherche par l'étude de la similarité globale entre les phrases, et entre les syllabes extraites.

3.1 Tâche de classification libre

Nous demandons aux auditeurs de juger la similarité globale entre échantillons sans leur donner de directives de stratégie ou de critères à privilégier. Les auditeurs devaient réaliser une classification libre des échantillons, à l'aide d'un logiciel créé pour l'expérience. D'abord, les auditeurs classaient les syllabes " to ", puis dans une deuxième étape les phrases. Une fois la classification réalisée, le logiciel demandait la stratégie globale suivie, puis, pour chaque classe réalisée, il était demandé une liste de qualificatifs, indiquant ce qui distinguait cette classe de sons par rapport aux autres. Cette tâche de verbalisation libre permet une compréhension des stratégies des auditeurs.

3.2 Création d'une distance holistique

A partir des classifications faites par les auditeurs, nous avons créé une distance holistique (ou de similarité globale), en comptant, pour chaque couple de sons et pour chacune des classifications faites par les auditeurs, le nombre de fois où les sons apparaissent dans des classes différentes : on obtient une matrice vérifiant toutes les propriétés d'une distance. Pour comprendre les éléments de cette distance, nous avons réalisé une analyse multidimensionnelle sur les distances entre phrases et entre syllabes (**Figure 2**). Pour expliquer les dimensions obtenues, nous avons fait des corrélations avec des mesures acoustiques (F0, jitter, pente spectrale, puissance, rapport énergie harmonique/bruit…), les informations sur le locuteur (âge et genre) et avec les critères perceptifs de l'expérience décrite au § 4.

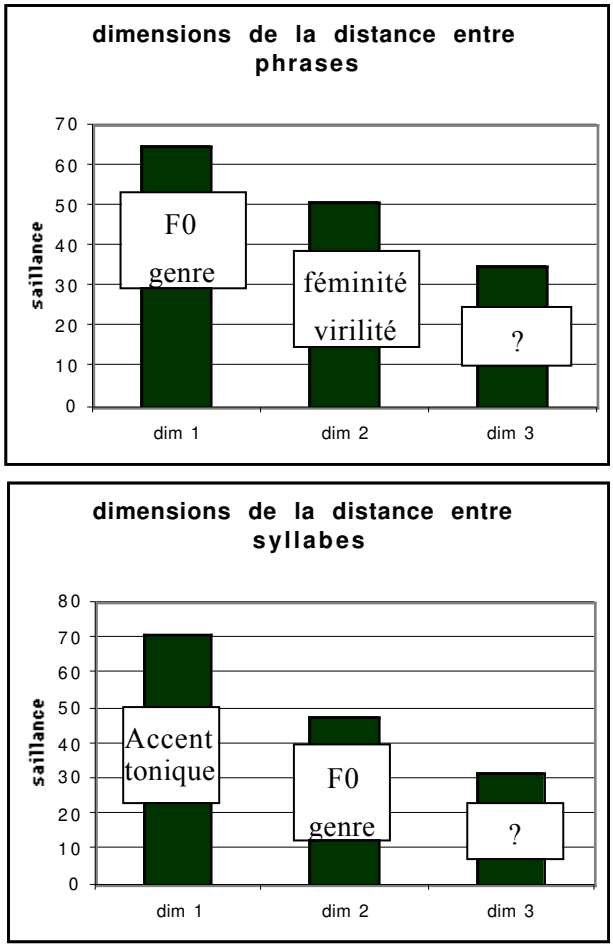

Figure 2 Les dimensions provenant de l'analyse multidimensionnelle (INDSCAL) pour la distance holistique entre syllabes et entre les phrases.

La hauteur (et le genre perçu qui lui est lié) est un critère très saillant, ce qui se retrouve dans les différentes expériences décrites dans la littérature ([Mur78, 80], [Kre96]). Nous constatons surtout que pour les syllabes, le critère de la position dans la phrase (position

tonique ou finale) est plus saillant que la hauteur. Donc les syllabes provenant du même locuteur (et donc avec des accents toniques différents) ont été classées à part : *la position dans la phrase est plus saillante perceptivement que l'individualité vocale du locuteur.*

3.3 Résultats de la verbalisation

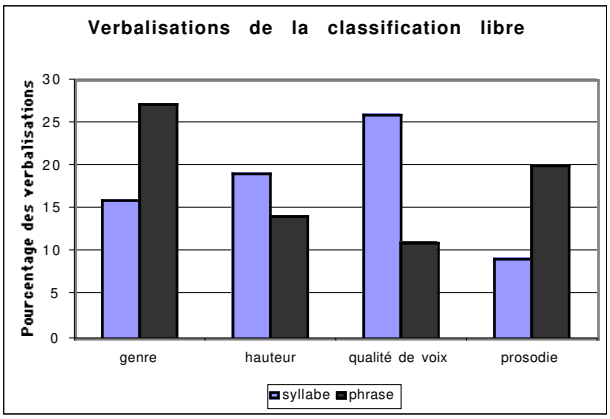

Figure 3 Pourcentages des types de qualificatifs servant à décrire les classes constituées par les auditeurs, pour les syllabes et les phrases.

Nous avons classé les qualificatifs que donnaient les auditeurs pour décrire les classes qu'ils avaient constituées en plusieurs catégories (genre, hauteur, qualité de voix, prosodie, autres). Nous remarquons dans la **Figure 3** que les auditeurs n'ont pas utilisé les mêmes critères : ils ont privilégié le critère du genre pour la phrase, alors que la hauteur était plus saillante pour la syllabe. En effet, on peut supposer que le matériau long permet de se faire une image du locuteur, et donc du genre. La prosodie était plus utilisée pour la phrase, car la syllabe est un support trop court pour des motifs prosodiques. *Pour un ensemble de locuteurs donnés, nous n'avons pas les mêmes critères de classification pour les syllabes et les phrases.*

4 Expérience d'écoute par axes prédéfinis

Pour mieux comprendre les différences de perception entre le matériau long et le matériau court, nous avons réalisé une expérience supplémentaire, où les auditeurs devaient juger le même matériau que précédemment (syllabes " to " et phrases) mais cette fois-ci en utilisant des critères prédéfinis, évalués sur des échelles graduées de 1 à 7.

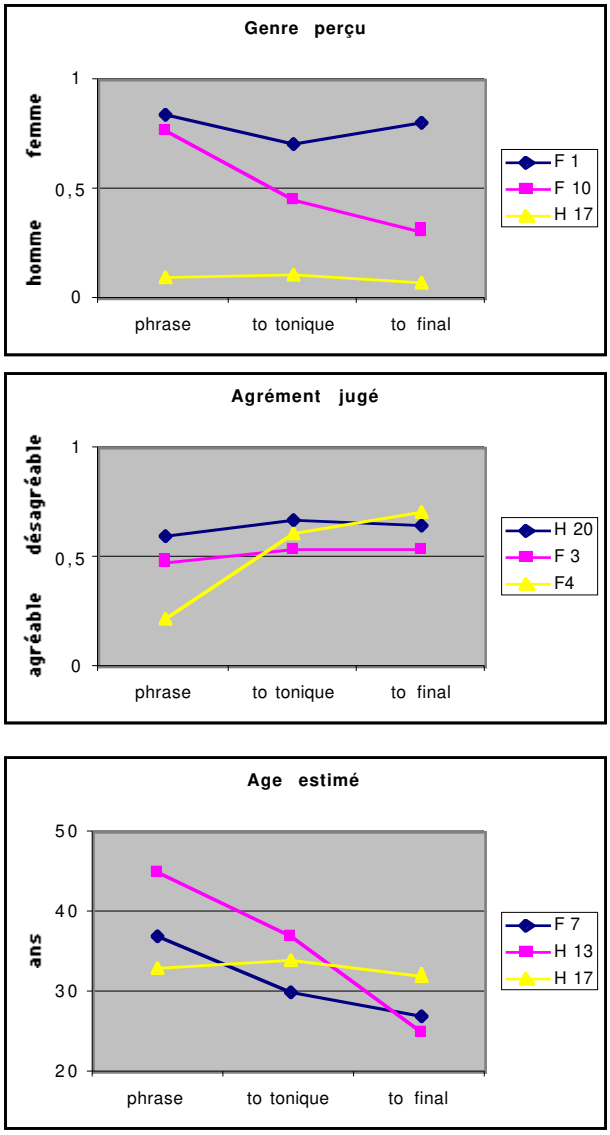

Figure 4 Moyenne des évaluations des auditeurs pour une phrase d'un locuteur donné, et pour les syllabes toniques et finale de la même phrase.

Les critères utilisés provenaient des verbalisations libres de l'expérience d'écoute holistique, que nous avons complétés par des critères couramment utilisés dans la littérature. Ces critères étaient : *agréable - désagréable, homme - femme, rapide - lent, avec - sans énergie, nasal - non nasal, avec - sans souffle, voilé - clair, tendu - détendu, puissant - faible, bonne - mauvaise prononciation, vulgaire-distingué, sans prétention-prétentieux, sympathique-antipathique*. Quand les auditeurs

avaient choisi le genre du locuteur, ils jugeaient les axes *grave-aigu, viril-peu viril, féminin-peu féminin* (pour un homme ou une femme). L'*âge minimum* et l'*âge maximum* étaient estimés en années.

Nous avons fait les moyennes des réponses données, et nous constatons que des différences sensibles peuvent apparaître entre les estimations faites pour les phrases et pour les différentes syllabes. Par exemple, il est bien connu que la prosodie entraîne des variations dans la hauteur. Cependant, nous remarquons dans la **Figure 4** que quand les auditeurs jugent le genre, ils peuvent également donner des estimations opposées pour la phrase en entier, ou pour une syllabe issue de cette phrase, comme c'est le cas pour la locutrice F10 (la phrase est perçue comme venant d'une femme, mais les syllabes comme venant d'un homme). Nous trouvons le même phénomène pour la perception de l'âge : par exemple le locuteur H17 a un âge estimé dans la trentaine pour la phrase et les syllabes, mais le locuteur H13 a un âge estimé dans la quarantaine pour la phrase, et dans la vingtaine pour la syllabe finale.

5 DEUX NIVEAUX D'ÉCOUTE : TIMBRE GLOBAL ET TIMBRE LOCAL

Les expériences que nous avons décrites mettent en évidence que les écoutes sont différentes pour un matériau court (timbre global) et un matériau bref (timbre local) :

A l'intérieur d'un timbre global (phrase) nous pouvons avoir de larges variations de timbre local (syllabes), comme le montrent les différences de perception de genre, âge, hauteur, et agrément entre les estimations des phrases et des syllabes extraites dans la **Figure 4**. Cependant nous n'avons pas obtenu un modèle prédisant la perception globale à partir des perceptions locales ;

Les variations de timbre local à l'intérieur d'une phrase sont perceptivement plus saillantes que les caractéristiques du locuteur, car les syllabes du même locuteur sont classées à part si elles n'ont pas la même position dans la phrase.

Par exemple, quand nous jugeons la force de voix d'une phrase donnée, une sorte de moyennage des différentes forces de voix au niveau syllabique a lieu. Si nous isolons des syllabes dans cette phrase, une syllabe tonique, elle aura une force de voix plus importante que la moyenne de la phrase.

6 DU TIMBRE À LA COHÉRENCE DU TIMBRE

Nous avons montré qu'il y a des variations des caractéristiques sonores du locuteur à l'intérieur même d'une phrase prononcée par un locuteur réel, mais nous n'avons pas de modèle des variations acceptables. Nous introduisons la notion de cohérence du timbre : cette notion de cohérence cherche à définir les limites que doivent respecter toutes les composantes d'une élocution, pour qu'elle semble avoir été prononcée par une seule personne réelle.

7 EXPÉRIENCE DE MONTAGE

7.1 Contraintes pour la substitution

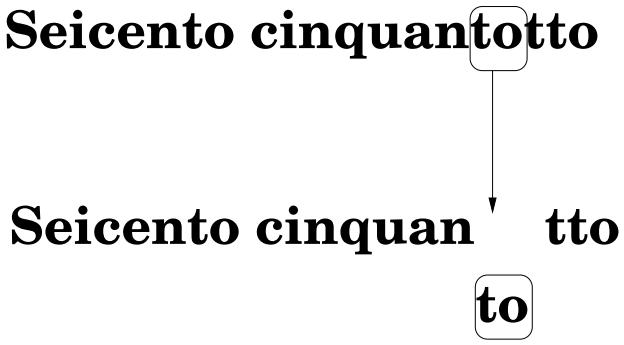

Figure 5 Le montage est fait en sélectionnant la syllabe " to " tonique dans une phrase, et en l'introduisant dans une autre phrase en remplacement de la syllabe tonique équivalente. Le montage est fait sans transformation supplémentaire du son.

Dans notre expérience de substitution de syllabes nous cherchons à déterminer les paramètres de la cohérence du timbre du locuteur. Cependant, avant de traiter la cohérence du timbre, il y a d'autres contraintes de source, comme le montrent les expériences dans le domaine de l'analyse de scènes auditives (Bregman, [Bre90]).

Contraintes d'environnement : si les segments que l'on concatène ont été enregistrés dans des conditions différentes (bruit ambiant, distance au microphone, système d'enregistrement), on percevra qu'il y a un montage. Le matériau de notre base est enregistré dans les mêmes conditions.

Contraintes de voisement : on ne doit pas avoir de changements abrupts dans les courbes de F0, ni dans les courbes de formants. Pour éviter ce genre de discontinuités, nous avons choisi de prendre une syllabe " bien détachée ", encadrée par deux consonnes plosives sourdes : il s'agit de la syllabe " to " tonique dans la phrase " seicento cinquantotto ", comme illustré dans la **Figure 5**.

Contraintes de la langue : ceci comprend le contenu linguistique et contenu prosodique. C'est pourquoi nous avons choisi de ne remplacer des syllabes que par d'autres syllabes ayant la même position dans une phrase équivalente (voir **Figure 5**).

Nous voyons donc que la cohérence du timbre du locuteur n'est qu'une contrainte qui s'ajoute aux autres contraintes : si le montage est accepté, on peut dire qu'il y a cohérence du timbre, mais dans le cas contraire, il se peut qu'une autre des contraintes du montage soit violée.

7.2 Conditions de l'expérience

Nous avons utilisé 20 locuteurs de la base EUROM pour l'italien, ce sont les mêmes locuteurs que nous avions étudiés dans une expérience préalable de classification libre et d'écoute par axes. Nous avons fait toutes les combinaisons des syllabes " to " toniques et des phrases, ce qui résulte en 380 montages et 20 phrases non montées. Les sons ont été présentés sur un questionnaire internet (www.limsi.fr/Individu/blas/experiences.html) par pages indépendantes de 40 sons, avec pour chaque page 20 phrases montées et les 20 phrases non montées, ce qui représentait 760 écoutes. Nous avons retenu 22 auditeurs, dont 13 ont écouté les sons avec des haut-parleurs dans une cabine isolée, et 9 ont utilisé des casques dans un bureau silencieux.

7.3 Tâche

Pour chaque élocution, les auditeurs indiquaient s'ils entendaient un locuteur (acceptation) ou un montage (rejet). On a obtenu 22 réponses par montage. A partir des réponses, on calculait une distance de montage, en faisant la moyenne des rejets sur le nombre total de réponses obtenues : une distance proche de 1 indique un montage majoritairement rejeté, et pour les montages acceptés la distance est proche de 0.

8 MONTAGE : RÉSULTATS

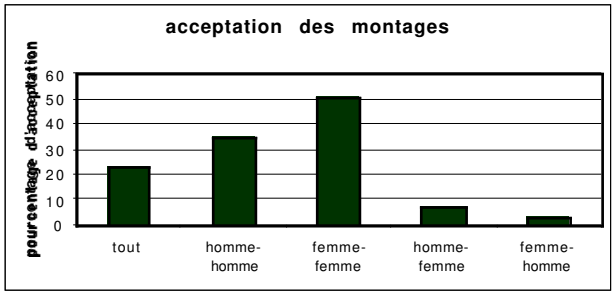

Figure 6 : L'acceptation des montages en fonction du genre du locuteur de la phrase porteuse et de la syllabe remplaçante.

Nous avons posé comme critère d'acceptation d'un montage le fait que la distance de montage soit inférieure à 0,5 (c'est-à-dire que la majorité des auditeurs l'ont accepté comme étant une phrase non montée). Nous pouvons voir dans la **Figure 6** que 23% des montages sont acceptés, et que ce pourcentage augmente si on se restreint aux montages faits entre locuteurs de même genre : la moitié des montages entre voix de femmes sont acceptés.

L'expérience de montage montre qu'une syllabe, provenant d'un locuteur donné, peut en fait être insérée dans des phrases provenant de locuteurs différents, résultant en une nouvelle phrase montée acceptable. Pour mieux comprendre les conditions d'acceptation nous allons procéder à une analyse de la distance perceptive de montage.

8.1 Dimensions de la distance de montage

La première dimension obtenue par analyse multidimensionnelle, qui est de loin la plus saillante, est liée au genre et à F0. La deuxième dimension est également liée à F0, mais de façon inverse pour les hommes (virilité estimée dans l'expérience d'écoute par axes, §4) et pour les femmes (féminité). La hauteur à elle seule explique la plupart de la variance de la distance de montage (84% de la variance pour les hommes et 77% pour les femmes). Nous observons donc que la qualité de voix (donc le timbre) revêt peu d'importance, et qu'au contraire la première dimension est la hauteur. On peut rapprocher ces résultats des recherches

faites en reconnaissance et discrimination du locuteur.

8.2 Prosodie et distance de montage

On peut distinguer deux effets dans le rejet dû à la hauteur :

1. Si la syllabe introduite est à une hauteur qui est très différente de la hauteur moyenne d'un locuteur, on pourra penser qu'il y a rupture de la cohérence du locuteur (cette syllabe est impossible pour le locuteur).
2. Si par contre la syllabe introduite reste dans l'ambitus habituel du locuteur, on peut penser qu'il y a rupture de la prosodie : c'est-à-dire que le locuteur peut émettre cette syllabe, mais que le contexte de la phrase supposerait une autre hauteur. Pour mieux comprendre les effets de la prosodie, nous avons tracé les contours mélodiques des différents montages, comme illustré dans la **Figure 7**.

Nous pouvons voir que toutes les syllabes dont les hauteurs sont situées autour d'une certaine valeur sont acceptées, et les autres rejetées : on peut alors définir une zone d'acceptation. Pour qu'une syllabe introduite soit acceptée, il n'est pas nécessaire qu'elle ait une valeur précise de hauteur, on peut plutôt dire que certaines valeurs sont plus probables que d'autres en fonction du contexte prosodique.

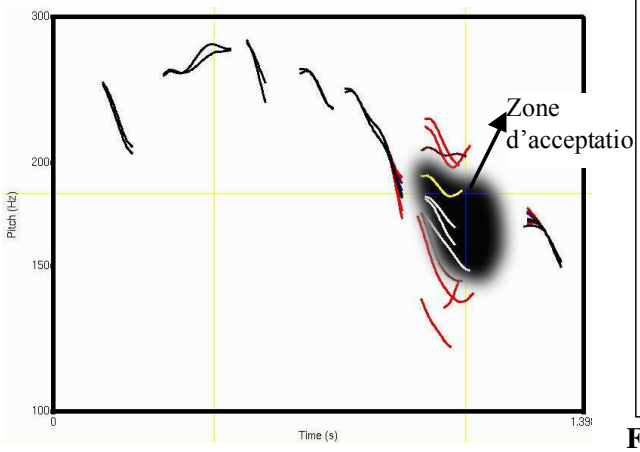

Figure 7 : Courbes mélodiques faites à partir des différents montages sur la phrase 6. On peut définir une zone d'acceptation, dans laquelle toutes les syllabes sont acceptées.

9 MONTAGE ET MODIFICATION DE F0

Nous avons établi que les syllabes qui sont dans une " zone d'acceptation " donnent des montages majoritairement acceptés, et les autres sont rejetées. Nous voulons tester si cette condition est suffisante, i.e. si on ramène une syllabe dans l'intervalle d'acceptation va-t-elle être acceptée ? et question inverse, si nous extrayons une syllabe acceptée de l'intervalle, va-t-elle être rejetée ?

Pour cela, nous avons fait une expérience complémentaire de montage, avec transformation de F0 avec la méthode PSOLA. Le matériau de cette expérience était :

> Une phrase porteuse (celle de la locutrice 6) avec les syllabes allogènes non modifiées, comme précédemment (19 montages et l'original)
>
> La phrase porteuse et syllabes allogènes modifiées de façon à avoir le même F0 que la syllabe remplacée (19 montages)
>
> La phrase porteuse avec syllabe originale modifiée en F0 (10 degrés de modification)

6 auditeurs ont participé à cette expérience, en utilisant des casques dans une salle silencieuse. Les 50 sons étaient présentés aléatoirement dans un questionnaire internet.

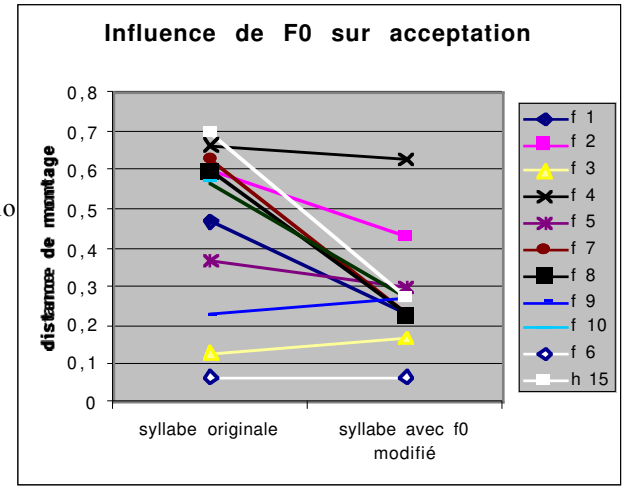

Figure 8 : L'influence de F0 dans l'acceptation des montages : à gauche nous avons la distance de montage pour la syllabe sans transformation, à droite la distance de montage pour la même syllabe transformée en F0 de façon à être dans l'intervalle d'acceptation.

Dans la **Figure 8**, nous voyons que quand les syllabes sont transformées en F0 de façon à entrer dans l'intervalle d'acceptation, l'acceptation du montage s'améliore. Cependant certains montages sont rejetés quand ils proviennent de locuteurs avec des qualités de voix nettement différentes : par exemple, dans la **Figure 9**, les locutrices 2 et 6 ne se mélangeaient pas car elles avaient des âges et féminités perçus comme trop différents.

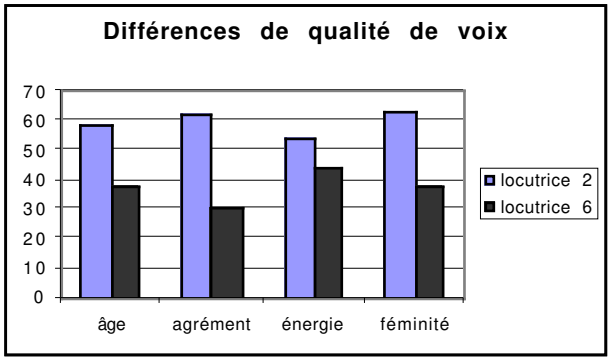

Figure 9 Les différences de qualité vocal perçues entre les locutrices 2 et 6 font que les montages sont rejetés, même quand le contour mélodique est respecté.

10 Conclusions des expériences de montage

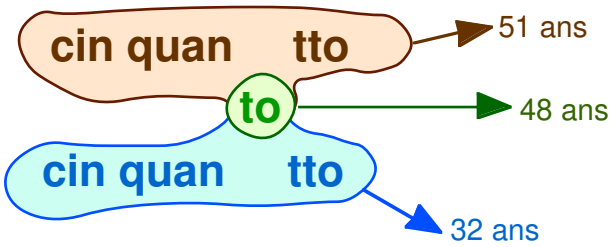

Figure 10 Un timbre local (syllabe) peut appartenir à deux timbres globaux différents. Ici deux phrases d'hommes d'âges perçus différents partagent une syllabe.

Les expériences de montage montrent qu'un timbre local (syllabe) peut être partagé par plusieurs timbres globaux comme résumé dans la **Figure 10** : l'individualité du locuteur ne repose pas sur la description instantanée du signal de parole.

11 Discussion

Les résultats de la voix parlée ne peuvent pas être directement transposés à la perception du timbre en général, car l'humain ne cherche pas les mêmes informations quand il entend une voix, ou quand il s'agit d'un son abstrait. Nous ne proposons donc pas de nouveaux paramètres du timbre en général, mais plutôt des méthodes de recherche.

La première conclusion générale de nos expériences de perception de la voix parlée porte sur la perception globale et locale. Selon que nous écoutons une phrase ou une syllabe détachée, nous n'utilisons pas les mêmes critères d'écoute. Nous avons constaté que, quand on demande aux auditeurs de juger selon les mêmes critères, la perception globale était différente de la moyenne des perceptions locales. Des variations importantes des valeurs perçues peuvent apparaître pour des syllabes détachées de leur contexte, par rapport à la phrase entière.

La deuxième conclusion porte sur la définition de la cohérence de la source : pour une phrase donnée, et une position dans cette phrase, plusieurs syllabes peuvent être utilisées (même si elles proviennent de locuteurs différents), à condition que les contours mélodiques soient respectés, et que les caractéristiques des locuteurs ne soient pas trop importantes.

Si on extrapole ces résultats à la perception du timbre en général, on peut dire que le fait de ne prendre qu'une note calibrée par instrument, et de comparer les notes pour plusieurs instruments, ne va donner qu'une partie de l'espace du timbre. Des paramètres perceptivement plus saillants du timbre peuvent apparaître en prenant divers échantillons du même instrument avec des modes d'émission, avec des tensions légèrement différents. En prenant des échantillons isolés, on risque également de négliger des paramètres importants dans l'écoute réelle musicale, puisque la perception globale d'une phrase musicale peut être significativement différente de la perception de ses notes isolées. Si nous considérons une phrase jouée par un violon, qui globalement sera perçue piano et spiccato, et que nous extrayons des notes de cette phrase, nous aurons une perception locale différente selon l'endroit choisi : l'attaque pourra être plus abrupte et le son plus tendu si nous prenons une note accentuée, et éventuellement la note finale sera perçue comme plus molle et détendue.

Il reste beaucoup de recherche à faire si on veut établir un modèle prédictif de la perception globale en connaissant les timbres locaux des différentes parties de la phrase musicale. Ces différences de perception sont ignorées par la recherche traditionnelle du timbre, qui utilise un seul échantillon isolé par instrument, considérant que l'ensemble de l'information de timbre est ainsi saisi.

Le critère de la cohérence du timbre peut être étendu aux différentes recherches de perception musicale, en posant comme question : est-ce que deux sons semblent provenir de la même source sonore quand ils sont concaténés ? Cette méthode de recherche peut être utile dans le développement d'un langage électroacoustique basé sur le timbre (ce qui serait une approche complémentaire des intervalles de timbre de McAdams [McA99]) : en effet, il est important, pour le compositeur, de savoir quand il transforme les sons, les transformations qui sont perçues comme une variation du matériau, et celles qui entraînent une rupture de la source, donc l'apparition d'un nouveau matériau dans l'œuvre.

12 BIBLIOGRAPHIE

[Bra99] Braun, A. et Cerrato, L. (1999), " Estimating speaker age accros languages ", XIV ICPhS, pp. 1369-1372.

[Bre90] Bregman, A.S. (1990), " Auditory scene analysis ", The MIT Press.

[Cas94] Castellengo, M. et Roubeau, B. (1994) "La notion de registre vocal", Rapport interne, LAM, ParisVI.

[Gre75] Grey, J. (1975), "An exploration of musical timbre", Mémoire de thèse, Université de Stanford.

[Kla90] Klatt, D.H et Klatt, L.C. (1990), " Analysis, synthesis, and perception of voice quality variations among female and male talkers ", JASA, Vol 87 (2).

[Kre96] Kreiman, J. et Gerratt, B.R. (1996), " The perceptual structure of pathologic voice quality ", JASA, Vol 100 (3).

[Lié99] Liénard, J.-S. et di Benedetto, M.G. (1999), " Effect of vocal effort on spectral properties of vowels ", JASA, Vol 106, pp. 411-422.

[Mca99] McAdams, S. (1999), " Perspectives on the contribution of timbre to musical structure ", Computer Music Journal, Vol 23 (3).

[Mur78] Murry, T., Singh, S. et Sargent, M. (1978), "Multidimensional classification of normal voice qualities", JASA, 64 (1).

[Mur80] Murry, T., Singh, S. et Sargent, M. (1980), "Multidimensional analysis of male and female voices", JASA, 68 (5).

[Sch66] Schaeffer, P. (1966), " Traité des objets musicaux ", Editions du Seuil.

[Sch85] Schmidt-Nielsen, A. et Stern, K.R. (1985), " Identification of known voices as a function of familiarity and narow-band coding ", JASA, Vol 77 (2).

[Slu96] Sluijter, A.M. et van Heuven, V.J. (1996), " Spectral balance as an acoustic correlate of linguistic stress ", JASA, Vol 100, pp. 2471-2485.

[Voi64] Voiers (1964), " Perceptual bases of speaker identity ", JASA, Vol 36, pp 1065-1073.

[Wal78] Walden, B.E., Montgomery, A.A., Gibeily, G.J., Prosek, R.A. et Schwartz, D.M. (1978), " Correlates of psychological dimensions in talker similarity", JSHR, Vol 21, pp. 265-275.

[Wes79] Wessel, D., (1979) " Timbre Space as a Musical Control Structure ", Computer Music Journal Vol 3, pp.45-52.