Vers un Réflexe Visio-Phonatoire

Robert Ruiz[a], Rui Da Silva Neves[b] , Clara Martinot[c] et Stéphane Vautier[d]

[a] LA.R.A, [b] L.T.C, [c] L.D.C.C, [d] C.E.R.P.P, Université de Toulouse II, 5 all. A.Machado, 31058 Toulouse cedex 1,France
courrier électronique: robert.ruiz@univ-tlse2.fr

RESUME
L'objet de la recherche est de mettre en évidence l'existence d'une possible influence de l'émotion provoquée par des images sur la phonation. Les images sélectionnées possèdent une "note émotionnelle" qu'il s'agit de corréler avec des caractéristiques acoustiques vocales. Ces dernières sont toutes issues du plan dynamique et sont appliquées à l'analyse de la voix d'un locuteur au comportement émotif. Les premiers résultats conduisent à sélectionner deux caractéristiques probablement sensibles à l'émotion.

I. INTRODUCTION

De nombreuses études attestent du contenu non linguistique de la parole. Ainsi l'état émotionnel d'un individu peut apparaître dans son expression vocale et éventuellement être détecté par des auditeurs. Les recherches dans le domaine ont eu pour support les voix de pilotes d'avion en état de stress du à des dysfonctionnements à bords [1], les voix de sujets chargés d'effectuer des tâches psychomotrices de difficulté variables [2], des tests de stress [1] etc … [3].
Ici les sujets ont à s'exprimer après avoir observé des images au contenu émotionnel élevé dont l'impact est connu et quantifié par une note de "plaisir" et une note "d'excitation" [4]. Les conditions expérimentales permettent de contrôler le vocabulaire et le déroulement chronologique des images. L'état initial d'anxiété du sujet est estimé par des tests psychologiques et son comportement est observé et analysé durant l'expérimentation.
De précédentes analyses vocales sur le même corpus n'ont pas permis d'observer des variations sensibles des caractéristiques acoustiques mesurées. Il s'agissait de la fréquence fondamentale moyenne, de l'écart-type associé et du jitter [5]. Peu exploré par les travaux antérieurs, le plan dynamique fait l'objet de cette étude et conduit à la mesure de caractéristiques temporelles et énergétiques.
La réaction aux stimuli émotionnels est apparue très variable d'un individu à l'autre conduisant à une analyse des manifestations vocales intra- plutôt qu'inter-individuelle. Le contexte visuel de l'expérimentation renforce ce point de vue. Il est courant de remarquer les grandes différences du comportement des spectateurs devant les images "choc". En conséquence, pour cette publication, les résultats ne concernent qu'un seul sujet.

II. METHODE EXPERIMENTALE

II.1. Images, Vocabulaire, Locuteur

Les images proviennent de la banque d'images I.A.P.S [4]. Chacune d'entre elles possède deux notes comprises entre 0 et 10 qui traduisent leur impact émotionnel en termes de valence affective du stimulus ("*pleasure*" *:* A*)* et d'excitation induite par ce stimulus ("*arousal*" *:* P). 24 images ont été sélectionnées mais seulement 12 seront exploitées ici. Ce sont 6 images qui possèdent des notes élevées dans les deux dimensions précédentes notées A+/P+, et 6 images appartenant au groupe A+/P- qui ont une note P faible et une note A élevée. Les images A+/P+ sont à caractère érotique, et les images A+/P- présentent par exemple des scènes d'accidents avec des blessures corporelles apparentes.
Après l'apparition d'un point de fixation sur l'écran d'ordinateur, chaque image est examinée par le sujet pendant 6 secondes puis une phrase apparaît en surimpression pendant 4 secondes. Le sujet doit la prononcer après qu'un signal sonore ("bip") ait été diffusé (4 secondes après la disparition de la phrase à l'écran). La phrase commence toujours par une structure Consonne1-Voyelle-Consonne2 (C1VC2). Cette structure désigne toujours un élément de l'image. S'il s'agit de la photographie d'un visage mutilé, la phrase est: "Kut est ensanglanté". Les structures CVC sont : [tip], [pit], [pak], [kap], [tyk], [kyt]. Les voyelles sont choisies pour couvrir un large domaine fréquentiel. Le choix des consonnes repose sur leur aptitude à être de nature impulsive afin de faciliter les analyses temporelles. Chaque structure C1VC2 est prononcée après deux images différentes appartenant chacune au groupe A+/P+ et au groupe A+/P-.
Le déroulement de l'expérience est entièrement automatisé. Le sujet gère lui-même le passage d'une image à la suivante en cliquant simplement sur la souris. Une phase d'entraînement est prévue pour que le locuteur se familiarise avec le procédé.
En préalable à l'expérience, deux test psychologiques sont effectués par les participants pour estimer leur état d'anxiété. Le locuteur (étudiant, 23ans) dont la voix est ici analysée, a été retenu pour ce premier dépouillement parce que les psychologues ont observé qu'il présentait une agitation comportementale attestant du trouble provoqué par les images. Les résultats aux tests préalables d'anxiété indiquent pourtant chez cette personne, un niveau d'anxiété moyen qui ne la prédispose donc pas à être un sujet sensible aux stimuli qui vont lui être proposés.

. Un programme spécifique développé dans l'environnement MATLAB permet d'estimer les valeurs numériques des caractéristiques.

III. CARACTÉRISTIQUES ACOUSTIQUES

Elles sont extraites du plan dynamique du signal, comme pour les précédentes mesures de période fondamentale et de jitter [5]. Ainsi, aucun pré-traitement n'est requis contrairement à la mesure des caractéristiques spectrales.

III.1. Durées

La détermination des durées de chaque phonème, et donc des durées de pause et de prononciation, est délicate, surtout lorsque l'échantillonnage rend possible la segmentation à l'échantillon près. Si des signes de l'émotion sont détectables par l'appareil auditif, il n'est pas improbable que les maxima d'amplitude du signal jouent un rôle substantiel dans cette perception. En effet, les durées des évènements sonores sont courtes et les effets de masque pré- et post-temporels se manifestent sans doute. Pour ces deux raisons, les durées mesurées dans les structures C1VC2 sont effectuées entre les maxima d'amplitude des phonèmes considérés.

Trois échantillons sont repérés. Ce sont les maxima d'amplitude de C1, de V et de C2. Deux durées sont calculées : d(C1V) et d(C1C2) qui sont respectivement la durée entre l'amplitude maximale de C1 et celle de V et la durée entre l'amplitude maximale de C1 et celle de C2.

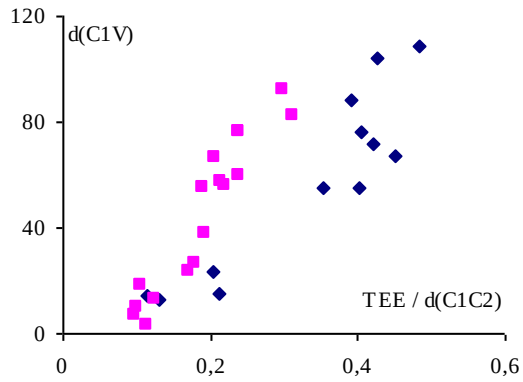
A ce stade exploratoire, le repérage n'est pas automatisé. Mais le découpage temporel basé sur la recherche des maxima d'amplitude devrait faciliter l'automatisation de la segmentation.

Le signal du "bip" étant parfaitement reproductible, il est facile de repérer toujours le même échantillon de début. Le temps de réaction tr est défini comme la durée entre l'échantillon choisi comme début du "bip" et l'amplitude maximale de C1.

III.2. Rapports d'Energie

Dans l'intervalle temporel d(C1C2) deux mesures de rapport d'énergie sont effectuées. La première est celle du rapport, noté Eav/Eap, entre l'énergie contenue dans d(C1V) et l'énergie contenue dans d(C2V).

La seconde consiste à définir une fenêtre temporelle de largeur égale à un échantillon et d'effectuer le rapport entre l'énergie du signal qu'elle contient et l'énergie contenue dans la fenêtre complémentaire à la durée d(C1C2). La largeur de la fenêtre initiale est incrémentée d'un échantillon entre chaque calcul de rapport d'énergie. L'objectif est de déterminer pour quel échantillon une égalité (ou quasi-égalité) énergétique existe entre les fenêtres antérieure et postérieure. Cet échantillon, converti en unité temporelle, est appelé Temps d'Equilibre Energétique (TEE). Cette mesure ne prend de sens qu'en proportion de la durée d(C1C2). C'est pourquoi, le rapport TEE / d(C1C2) est calculé.



IV. RESULTATS

Tableau 1. Mesures des caractéristiques acoustiques pour les prononciations "au repos". Les durées d(C1V) et d(C1C2) sont en millisecondes.

C1VC2	d(C1V)	d(C1C2)	Eav/Eap	TEE/d(C1C2)
Tuk	17,85	298,23	0,281	0,106
Tuk	76,08	331,95	0,790	0,239
Tuk	57,01	318,05	0,539	0,215
Tip	82,43	289,37	0,714	0,312
Tip	92,18	326,39	0,736	0,298
Kap	59,57	375,10	0,422	0,238
Kap	76,01	331,88	0,790	0,239
Pak	7,12	255,19	0,355	0,098
Pak	3,29	313,06	0,094	0,114
Pak	9,89	270,29	0,364	0,100
Kut	55,58	348,64	0.322	0,219
Kut	37,98	322,04	0,157	0,194
Kut	66,46	348,87	0,704	0,207
Kut	55,08	367,76	0,416	0,191
Pit	26,55	290,66	0,128	0,179
Pit	13,06	285,17	0,178	0,124
Pit	23,76	301,75	0,248	0,171

Le calcul des moyennes et écart-types des deux grandeurs énergétiques pour toutes les structures phonétiques confondues (Tableau 3) montre que les images provoquent une hausse sensible des valeurs numériques de ces caractéristiques.

Les représentations graphiques des figures 1 à 4 renforcent ce résultat. Elles montrent également que les associations graphiques de ces deux grandeurs avec la durée entre les maxima d'amplitude des deux consonnes peuvent contribuer à identifier les prononciations "au repos" de celles après la vue des images (figures 1 et 2). Cela n'est pas le cas avec la durée d(C1V) entre la première consonne et la voyelle (figures 3 et 4).

Tableau 2. Mesures des caractéristiques acoustiques pour les prononciations après la vue des images. Les durées d(C1V), d(C1C2) et tr sont en millisecondes.

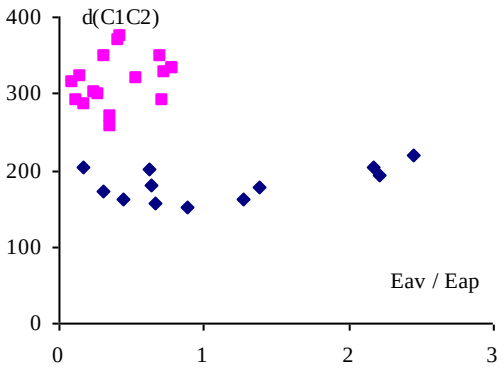
C1VC2 / Image	tr	d(C1V)	d(C1C2)	Eav/ Eap	TEE / d(C1C2)
Tuk /2	1068,1	55,28	203,36	0,162	0,401
Tip /5	795,10	13,99	157,57	0,659	0,113
Kap /7	958,98	55,35	202,47	0,627	0,353

Pak /9	1034,4	14,90	172,20	0,308	0,211
Kut /12	1259,7	67,17	152,13	0,880	0,451
Pit /19	791,32	23,13	162,34	0,442	0,203
Pit /4	906,69	71,72	162,77	1,266	0,420
Pak /8	991,16	12,65	181,13	0,631	0,131
Kut /13	820,98	104,51	218,75	2,444	0,425
Tuk /17	896,44	75,87	177,14	1,385	0,405
Pit /21	806,44	87,98	194,54	2,206	0,392
Tip /23	900,95	108,73	204,60	2,167	0,483

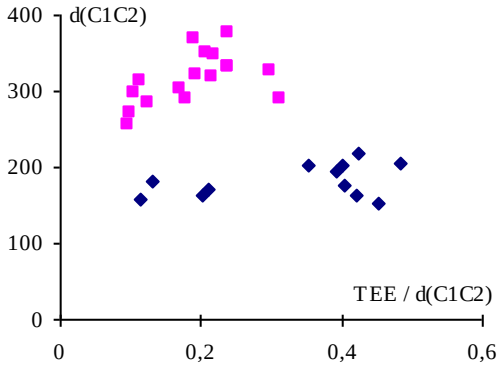
Tableau 3. Comparaison des moyennes (écart-types entre parenthèses) sans voir ("repos") et après la vue des images ("images") pour toutes les structures C1VC2.

	"repos" (n=17)	**"images" (n=12)**
Eav / Eap	0,426 (0,242)	1,098 (0,792)
TEE / (C2-C1)	0,191 (0,066)	0,332 (0,130)

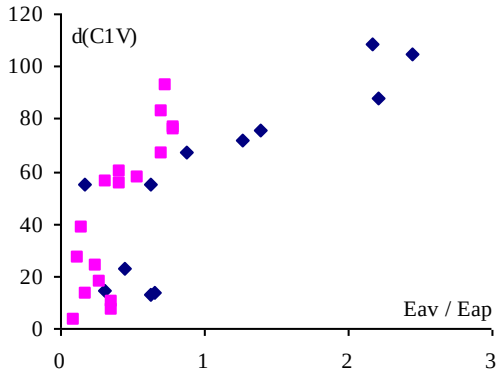
Figures 1 à 4. Valeurs numériques conjointes des caractéristiques acoustiques mesurées pour les prononciations au "repos" (carrés) et celles après les images (losanges).



Les images 2, 5, 7, 9, 12, 19 appartiennent au groupe A+/P+ et les images 4, 8, 13, 17, 21, 23 au groupe A+/P- (Tableau 2). Les deux ensembles bien distincts de "losanges" sur la figure 1 ne suivent pas exactement le même découpage. Mais l'ensemble le plus à droite comprend 5 des 6 images A+/P-. De même, 6 des 7 "losanges" les plus à gauche sur la figure 1 correspondent à des images du groupe A+/P+.
L'émotion provoquée par les images entraîne, chez ce sujet:
- pour toutes les images, une diminution sensible de la durée d(C1C2) conjointement avec
- une augmentation de TEE/d(C1C2) et de Eav/Eap pour une partie seulement des images (respectivement 8 et 5 images sur 12).



Il est à noter qu'aucune variation significative de la fréquence fondamentale moyenne, de l'écart-type associé et du jitter n'a pu être observée lors des comparaisons entre voyelles issues des même C1VC2 pour les prononciations au repos et après les images.
Les valeurs numériques des temps de réaction tr ne permettent pas d'identifier deux groupes en relation avec les types d'image A+/P+ et A+/P-. Elles devront être comparés avec ceux obtenus après la diffusion d'images au contenu émotionnel neutre.
L'analyse des résultats pour la durée d(C1V)



conjointement avec les deux grandeurs énergétiques ne permet pas d'isoler des tendances aussi nettement que pour la durée d(C1C2) (figures 3 et 4). Cependant, l'analyse individuelle de la durée d(C1V) révèle comme pour d(C1C2) des variations numériques sensibles (Tableau 4).

Tableau 4. Durées moyennes et écart-types associés (entre parenthèses) mesurés sur les structures C1VC2 sans voir les images ("repos"), après la vue de toutes les images ("images") et après la vue uniquement de celles appartenant aux groupes A+/P+ et A+/P-.

	d(C1V) ms	**d(C1C2) ms**
"repos" (n=17)	44,70 (29,14)	316,14 (33,16)
"images" (n=12)	57,61 (34,78)	182,42 (21,86)
A+/P+ (n=6)	38,30 (23,58)	175,01 (22,60)
A+/P- (n=6)	76,91 (34,80)	189,82 (20,23)

Après la vue des images on constate une diminution de la durée entre les maxima des consonnes dont une cause possible est la diminution de l'intervalle temporel entre le maxima de la voyelle centrale et la consonne postérieure. En effet, la durée d(C1V) entre le maxima de la consonne antérieure et celui de la voyelle augmente. Enfin, la durée d(C1V) moyenne analysée par groupes d'images A+/P+ et A+/P- varie du simple au double ce qui révèle peut-être une aptitude favorable à isoler les influences, comme pour la représentation conjointe de d(C1C2) avec Eav/Eap (figure 1).

V. DISCUSSION ET PERSPECTIVES

La méthode expérimentale mise en place présuppose que l'émotion provoquée sera "verbalisée". L'objectif est de le démontrer. Pour cela, le critère de choix du premier sujet a été basé sur l'observation comportementale des psychologues. Or, il est possible que les signes d'agitation comportementale soient un moyen d'évacuer le trop plein d'émotion ressentie et que celle-ci ait disparu lors de la prononciation. Il importe donc de compléter les analyses par l'examen des voix de quelques sujets impassibles et de conforter les résultats pour les sujets présentant des signes extérieurs de trouble.

Pour chacun d'eux, il conviendra également d'effectuer les mesures avec des images au contenu émotionnel neutre. Ainsi l'effet de l'image pourra être estimé. La situation non émotionnelle est ici une situation où le locuteur ne voit pas d'image mais seulement un texte où figure la liste des structures phonétiques qu'il doit prononcer.

Par ailleurs, la méthode utilisée présuppose que l'expression vocale de l'émotion se manifeste sur le premier mot à prononcer. Or l'instant de début des manifestations (acoustiques) émotionnelles, leur intensité, leur durée ne sont pas connus.

D'un point de vue acoustique, les résultats montrent que même si les caractéristiques dérivées de la fréquence fondamentale (dont les variations sont classiquement associées à l'expression vocale de l'émotion) ne sont pas sensibles ici aux facteurs étudiés, il est possible que d'autres grandeurs le soient, par exemple celles issues du plan dynamique, en lien direct avec le débit et l'enveloppe temporelle du signal. Le plan spectral reste par ailleurs à examiner.

Le vocabulaire utilisé pour l'étude est restreint. Le choix des consonnes limite la généralisation des résultats même si ceux-ci sont confirmés ultérieurement par les analyses des voix d'autres locuteurs. D'autre part, la nature impulsive du signal des consonnes choisies, bien que différentes, est propice au repérage des maxima d'amplitude. Il conviendra de tester et, le cas échéant, d'adapter les caractéristiques aux autres consonnes.

Une éventuelle détection de l'émotion sur du vocabulaire courant nécessite des critères adaptables à tout type de structure phonétique complexe. Les caractéristiques énergético-temporelles sont peut-être mieux adaptées que leurs homologues énergético-fréquentielles.

VI. CONCLUSION

Les critères proposés pour détecter l'émotion d'origine visuelle sont prometteurs. Les résultats révèlent leur aptitude à la détection mais également une certaine disposition à différencier le type d'image à l'origine de la manifestation émotionnelle. Leur mesure sur de nombreux locuteurs est nécessaire pour les classer réellement au rang d'indicateur potentiel, mais l'analyse des résultats doit demeurer intra-individuelle tant les réactions aux stimuli sont variables d'un locuteur à l'autre.

Le réflexe visio-phonatoire, ou comment la vision interagit de manière non consciente sur la phonation, peut exister dans la vie quotidienne dès lors qu'il s'agit par exemple de relater, de décrire une situation perturbante dont on a été un témoin oculaire. L'expérimentation de laboratoire qui a été réalisée pour en rechercher les indicateurs acoustiques semble pouvoir simuler les effets attendus et offre donc des perspectives de recherche étendues.

RÉFÉRENCES

1. R. Ruiz, E. Absil, B. Harmagnies, C. Legros et D. Poch, *Time- and Spectrum-Related Variabilities in Stressed Speech under Laboratory and Real Conditions*, Speech Com., **20** (1-2), 111-129 (1996).
2. G.R Griffin, C.E. Williams, *The Effects of Different Levels of Task complexity on Three Vocal Measures,* Aviat. Space Environ. Med., **58**, 1167-1170 (1987).
3. R. Ruiz, C. Legros, A. Guell, *Voice Analysis: Application to the Study of the Influence of a Workload,* J. Acoustique, **3**, 153-159 (1990).
4. Center for the Study of Emotion and Attention (CSEA-NIMH) International Affective Picture System: Digitized Photographs. Gainsville, FL: The Center for Research in Psychophysiology, University of Florida (1999).
5. R. Ruiz, R. Da Silva Neves, C. Martinot, S. Vautier, *Interactions between Emotional Content of Pictures and Acoustic Features of Speech,* 17^{ih} International Congress on Acoustics, Roma, Italy (2001).