Prosodie multimodale
Les enchères chantées aux Etats-Unis

Gaëlle Ferré
LLING, Chemin de la Censive du Tertre, BP 81227, 44312 Nantes, cedex 3
Gaelle.Ferre@univ-nantes.fr

RESUME

Cet article propose une analyse prosodique multimodale de la vente aux enchères chantée venant des Etats Unis. A partir d'un corpus d'enregistrements de 6 locuteurs et en nous appuyant sur l'analyse prosodique de Kuiper, K. & Tillis F. (1985), nous essayons de voir comment les gestes, nécessaires dans ce type d'interaction, sont alignés avec la parole alors que le débit des locuteurs est très rapide et que le contenu verbal est contraint par la structure rythmique du chant.

ABSTRACT

Multimodal Prosody. The auction chant in the United States

This paper proposes a multimodal prosodic analysis of the auction chant that is practiced in the US. Drawing upon a corpus that involves 6 auctioneers and relying upon the prosodic analysis of the chant by Kuiper, K. & Tillis F. (1985), we investigate how gestures, which are necessary in this type of interaction, align with speech despite a fast speech rate and the fact that verbal content is strongly constrained by the rhythmic structure of the chant.

MOTS-CLES : Communication multimodale, prosodie, vente aux enchères chantée.
KEYWORDS : Multimodal communication, prosody, auction chant.

1 Introduction

La vente aux enchères traditionnelle suppose une interaction entre un commissaire-priseur, chargé de la vente d'un produit, et un public d'acheteurs potentiels. Elle constitue un mode de communication multimodal par excellence dans la mesure où le commissaire-priseur, tout en annonçant le prix de l'enchère verbalement doit aussi désigner l'un des acheteurs potentiels dans le public, double action répétée jusqu'à la vente du produit. Dans le sud des Etats-Unis, est apparue une variante de l'enchère anglaise, dite « enchère ascendante », appelée « auction chant » et apparentée aux Negro spirituals puisque son origine remonte à la vente de tabac dans les plantations (Kuiper, K. and Tillis F., 1985). Une partie de l'enchère est chantée ou psalmodiée. L'enchère chantée a fait l'objet de deux études prosodiques dans (Kuiper, K. and Haggo D., 1984; Kuiper, K. and Tillis F., 1985). Kuiper propose par ailleurs (Kuiper, K., 1992, 2000) une analyse discursive des ventes aux enchères anglaises chantées ou parlées. Les travaux de Kuiper et ses collègues sont à notre connaissance les seuls travaux existants sur l'enchère chantée.

Par ailleurs, les gestes ont été analysés dans la vente aux enchères parlée dans Heath, C. and Luff P. (2007, 2011). Ils montrent que l'apogée du geste coïncide avec la syllabe

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 177–184,
Grenoble, 4 au 8 juin 2012.

nucléaire des incréments, sans pour autant proposer une analyse prosodique. Nous nous sommes donc interrogée sur l'alignement geste-voix dans le cas de l'enchère chantée : l'une des caractéristiques prosodiques de ce type d'enchère, nous allons le voir, est un débit de parole très rapide. Or, la différence de granularité du geste et de la parole rend l'alignement des deux modes difficile dans un tel contexte. Le locuteur (ici le commissaire-priseur) devra donc choisir une stratégie pour que ses gestes ne soient pas en complet décalage avec sa parole.

A partir d'un corpus d'enregistrements vidéo décrit dans la section 2, nous proposerons une analyse discursive et prosodique de la vente aux enchères chantée en nous appuyant sur et en complétant les travaux de Kuiper et ses collègues. Puis, nous analyserons la manière dont le geste s'aligne avec la parole.

2 Données et méthode

2.1 Enregistrements vidéos

Afin de pouvoir réaliser une analyse prosodique et multimodale de qualité, il nous a semblé impossible de travailler sur des fichiers vidéo de vente aux enchères chantées réalisés en contexte naturel beaucoup trop bruyants pour l'analyse acoustique de la parole. En revanche, ce type de vente fait aussi l'objet de concours enregistrés avec une qualité à la fois de l'image vidéo et du son qui permet une analyse multimodale. Les concours présentent également un avantage supplémentaire : le type de vente et la durée de chaque vente y sont réglementés ce qui permet une réelle comparaison entre des ventes réalisées par différents locuteurs. L'interaction n'en est pas pour autant artificielle puisque des ventes sont effectivement réalisées auprès d'un public, la somme récoltée étant ensuite reversée à une association. Les lots vendus sont des objets de la vie courante de valeur différente (bottes, téléphone portable, etc). C'est sur ce type de fichiers que repose cette analyse. Enfin, les fichiers choisis sont des enregistrements de finalistes du concours, ce qui garantit la qualité de la prestation. Le corpus compte 6 locuteurs (3 hommes, 3 femmes ; 1 fichier audio-vidéo par locuteur). Chaque locuteur, après une brève présentation, réalise 3 ventes. Chaque vente comporte elle-même une partie parlée et une partie chantée qui ont été distinguées au niveau de l'annotation et dont nous présenterons la structure dans l'analyse discursive.

2.2 Traitement des données

L'intégralité des ventes dure environ 20 minutes (environ 2.30 minutes par vente). Chaque fichier a été transcrit sous Praat avec un alignement sur le fichier audio séparé de la vidéo au niveau des mots (transcription orthographique) et des syllabes (transcrites en SAMPA). La transcription a été ensuite vérifiée par un locuteur natif de l'anglais. Les gestes manuels et leur apogée (point d'extension maximale) ont été annotés sous Elan. Pour les unités gestuelles, nous avons compté le début du geste sur l'image qui précède immédiatement le début du mouvement jusqu'à la fin de la rétraction du geste. Dans le cas où deux gestes s'enchaînent, nous avons compté le début du deuxième geste à partir du changement de direction de la main. Nous avons distingué trois types de gestes : les pointages simples (index ou main sur la tranche tendu(e) vers un membre du public), les pointages complexes qui sont bi-dimensionnels

(voir McNeill, 2005 ; par exemple, lorsque la main tendue vers un membre du public est en supination et comporte donc une part de geste métaphorique, ou lorsque le locuteur lève deux doigts vers un membre du public, où le pointage comporte une partie emblématique) et les autres types de gestes qui n'impliquent pas de pointage (dans distinguer entre les iconiques, les battements, les emblèmes et les métaphoriques). Dans les parties parlées, on dénombre 65 pointages simples et 79 autres types de gestes (nb total = 144). Dans les parties chantées, nous avons annoté 212 pointages simples, 243 pointages complexes et 59 autres types de gestes (nb total = 514).

Nous avons également noté sous Elan les différents actes dans la partie chantée de chaque vente après avoir importé les annotations Praat. Ils seront présentés dans l'analyse discursive.

2.3 Annotation discursive des ventes aux enchères chantées

Kuiper & Tillis (1985) ont proposé une analyse structurelle des ventes aux enchères anglaises. Celles-ci, selon les auteurs, suivent donc la structure suivante, structure qui se retrouve largement dans les ventes de notre corpus. Une vente aux enchères comprend 5 phases : la description du lot, la mise en vente (prix initial), l'enchère (phase qui comprend un certain nombre d'actes – les prix incrémentés, ainsi que d'éventuels énoncés annonçant la fin proche de la vente), la vente et l'épilogue (facultatif). Dans notre corpus, la vente est invariablement réalisée par l'énoncé « and I have sold it » qui correspond au *adjugé, vendu* du français. L'épilogue est toujours présent et consiste à redonner le prix de vente et à demander le numéro d'identification de l'acheteur.

Les deux auteurs ne précisent pas quelles parties de la vente aux enchères sont parlées et quelles parties sont chantées. Dans notre corpus, la partie chantée de chaque vente ne concerne que la mise en vente et l'enchère. Les autres phases sont parlées. Afin de conduire notre analyse multimodale, nous avons donc annoté les actes de langage produits dans ces parties chantées. Heath & Luff (2007), dans leur analyse du fonctionnement des ventes aux enchères, signalent que le commissaire-priseur trouve initialement deux (et uniquement deux) enchérisseurs dans le public. A chaque nouvel incrément, le commissaire-priseur annonce le prix et effectue un pointage vers l'un des deux enchérisseurs. Le but de ce pointage est de proposer à cet enchérisseur la nouvelle somme. Si cette somme est acceptée, il propose alors la somme incrémentée au deuxième enchérisseur etc. Ce qui est important dans notre corpus, c'est qu'en comparaison avec une vente aux enchères parlée, il n'y a pas de place pour les pauses et la réflexion, ainsi, la somme est répétée jusqu'à ce que la proposition soit acceptée par un enchérisseur. Ce n'est qu'une fois la proposition acceptée qu'il peut à nouveau incrémenter la somme et chercher un nouvel acquéreur. De plus, l'opposition entre deux enchérisseurs seulement est moins marquée que dans l'enchère traditionnelle et les gestes de pointage sont répétés également. Dans notre annotation, nous avons donc distingué différents actes dans l'enchère : la proposition (en distinguant entre l'énoncé de la somme et l'énoncé de l'incrément) et l'acceptation, puis la répétition de la proposition et la répétition de l'acceptation, et enfin les actes d'encouragement. Voici un extrait et son analyse :

(...) seven seventy five	ACCEPTATION	*sept (cent) soixante quinze*
no eight hundred for you (...)	PROPOSITION	*non huit cents pour vous*
now eight hundred dollar one time	RÉP. PROPOSITION	*et huit cents dollars une fois*
eight hundred dollar	RÉP. PROPOSITION	*huit cents dollars*
seven seventy five	RÉP. ACCEPTATION	*sept (cent) soixante quinze*
eight hundred	RÉP. PROPOSITION	*huit cents*
we got to go	ENCOURAGEMENT	*il faut y aller*
eight hundred dollar	RÉP. PROPOSITION	*huit cents dollars*

C'est sur la base de ces actes de langage qu'ont été réalisés les calculs prosodiques et les alignements gestuels présentés dans les sections suivantes.

3 Résultats

3.1 Analyse prosodique des ventes aux enchères chantées

Pour cette analyse, nous avons retenu la durée des pauses, la durée syllabique et phonémique, ainsi que F0. Nous avons préféré ne pas faire de calculs sur l'intensité car le type de micro utilisé par les locuteurs ne garantit pas une mesure stable de ce paramètre.

3.1.1 Durée

Afin d'analyser la prosodie de la partie chantée du corpus, nous avons comparé ses caractéristiques aux parties parlées. Dans leur article, Kuiper & Tillis (1985) signalent que la vente aux enchères chantée est plus rapide que la parole mais ils ne disent pas à quelle parole ils l'ont comparée ni avec quel outil d'analyse. Ils précisent également que cette perception d'un débit rapide est due à une plus grande fluidité (moins de pauses). Dans notre corpus, nous avons trouvé que les pauses sont légèrement moins nombreuses dans la partie chantée (1 pause toutes les 3.1 sec en moyenne) que dans la partie parlée (1 pause toutes les 2.8 sec en moyenne), mais l'écart entre les deux types de parole n'est pas significatif. Ce qui est significatif en revanche, d'après le test-t de Student que nous avons réalisé[1], est que la durée moyenne des pauses est significativement différente entre la parole et le chant ($t = -4.5791$, $df = 181.363$, p-value < 0.01) avec une durée moyenne de 0.216 sec pour le chant et 0.409 sec pour la parole. Cette différence s'explique par le fait que les pauses dans les parties chantées sont strictement respiratoires, contrairement aux parties parlées.

En ce qui concerne les syllabes, nous observons que la réduction syllabique est plus importante dans le chant que dans la parole (par exemple « seventy » *soixante-dix*, normalement prononcé /se.vən.ɾi/ en anglais américain, est régulièrement prononcé /sev.ni/ dans la partie chantée). Il en va de même pour la réduction phonétique (« five » *cinq*, normalement prononcé avec une diphtongue /faɪv/, est régulièrement prononcé avec une monophtongue /fav/ dans la partie chantée). Notre test statistique montre une différence de nombre de phonèmes prononcés par syllabe entre le chant et la parole ($t = -7.649$, $df = 3669.762$, p-value < 0.01) avec une moyenne de 2.4 pour

[1] Statistiques réalisées sous R.

le chant contre 2.6 pour la parole. Ceci explique que la durée moyenne des syllabes soit plus élevée dans la parole que dans le chant ($t = -26.1736$, $df = 3173.639$, p-value < 0.01) ainsi que la durée phonémique moyenne ($t = -27.2333$, $df = 3004.949$, p-value < 0.01). Il est possible cependant, que si nous avions pu comparer ces données à de la parole conversationnelle, l'écart de durée entre le chant et la parole n'ait pas été aussi élevé.

Enfin, pour chaque syllabe, nous avons attribué une valeur *longue* ou *brève* en divisant en deux la plage des durées. Ce critère assez grossier s'est révélé rendre parfaitement compte de la perception que l'on peut se faire de la durée de ces syllabes. Ces valeurs nous ont permis d'établir des schémas rythmiques pour tous les actes de langage rencontrés dans le chant. Comme nous le disions dans l'introduction, il s'agit plus d'une psalmodie que d'un chant car la variabilité rythmique est assez importante. Cependant, certains schémas sont plus fréquents que d'autres. Parmi les schémas les plus fréquents, on distingue deux groupes : dans le premier groupe, les actes de langage comprennent entre 2 et 7 syllabes, toutes brèves. Dans le deuxième groupe, les actes de langage comprennent entre 2 et 8 syllabes ; toutes sont brèves excepté la dernière qui est longue.

3.1.2 FO

En ce qui concerne la mélodie, Kuiper & Tillis (1985) distinguent deux modes dans la partie chantée : le mode « drone » dans lequel toutes les syllabes sont prononcées à la même hauteur mélodique, excepté la syllabe nucléaire de chaque groupe qui présente un mouvement mélodique descendant. Dans le mode « shout », c'est au contraire la première syllabe qui présente une mélodie plus élevée. Ceci correspond bien à ce que l'on observe sur notre corpus, même si nous n'avons pas refait ces calculs. En revanche, ils signalent également que les plages intonatives sont comprimées dans les deux modes. Notre corpus confirme ce résultat. Nous avons extrait la F0 dans Praat automatiquement toutes les 0.01 sec et avons comparé les résultats obtenus pour la parole et pour le chant en séparant les hommes et les femmes. Nous trouvons que la F0 est plus significativement plus élevée dans le chant que dans la parole pour les deux groupes de locuteurs (Hommes : $t = 16.2568$, $df = 16088.60$, p-value < 0.01 ; Femmes : $t = 13.3736$, $df = 17252.41$, p-value < 0.01). La moyenne de F0 est de 199.1 Hz pour les hommes et de 263.3 Hz pour les femmes dans le chant, contre 190.2 Hz pour les hommes et 252.3 Hz pour les femmes dans la parole. Nos résultats confirment aussi ceux de Kuiper & Tillis (1985) en ce qui concerne la compression de la plage intonative, avec cependant une légère différence entre les hommes et les femmes, ainsi que le montre la figure 1 qui représente l'histogramme des valeurs de F0 pour les deux groupes de locuteurs dans la parole et dans le chant.

Les histogrammes de la Figure 1 montrent que pour les hommes, la distribution des valeurs de F0 est normale dans la parole (1a), alors que l'on observe une distribution négativement asymétrique dans le chant (1b), avec un pic de valeurs comprises entre 175 et 200 Hz. Cette compression de la plage intonative est illustrée dans la Figure 2, qui affiche la courbe de F0 d'un locuteur masculin dans la partie chantée de l'enchère.

Pour les femmes, les histogrammes de la Figure 1 montrent que la distribution des valeurs de F0 est négativement asymétrique dans la parole (1c), mais qu'elles sont

présentes dans une plage assez étendue allant de 175 à 350 Hz, alors que pour les parties chantées, l'on observe une distribution bimodale avec un pic autour de 200 Hz et un deuxième autour de 250 Hz.

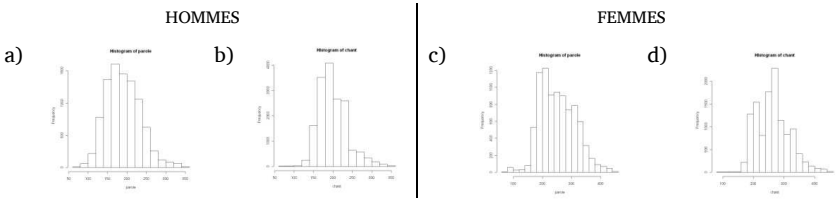

Figure 1 – Histogrammes des valeurs de F0 en Hz pour les hommes dans la parole (a), dans le chant (b), pour les femmes dans la parole (c), dans le chant (d).

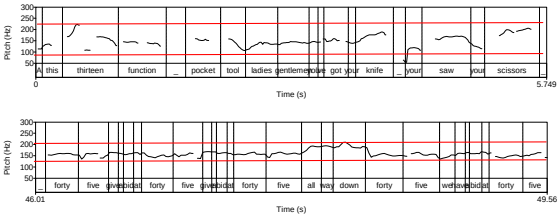

Figure 2 – Courbe de F0 en Hz d'un extrait de parole (en haut) et d'enchère chantée (en bas) pour le même locuteur masculin.

3.2 Analyse gestuelle des ventes aux enchères chantées

Comme le montrent Heath & Luff (2007), les différents actes émis dans la phase d'enchère d'une vente aux enchères traditionnelle sont accompagnés entre autres de pointages du commissaire-priseur vers un enchérisseur. Le rôle des pointages est de lui donner la possibilité d'accepter la somme incrémentée. Les auteurs montrent également que l'extension maximale (apogée) du pointage coïncide avec la syllabe nucléaire du groupe intonatif, même s'ils n'ont pas réalisé une étude prosodique de leur corpus. Or, dans notre corpus, l'on constate deux choses : le débit de parole extrêmement rapide de la partie chantée ne permet pas au locuteur de faire correspondre un geste à un acte de langage, car la granularité du geste est plus large que celle de la parole. Le locuteur va donc devoir trouver une stratégie pour aligner au mieux gestes et parole. L'une des stratégies possibles consiste à réduire la durée des gestes dans le chant par rapport à la parole en réduisant leur amplitude. Le test-t de Student montre cependant qu'il n'y a aucune différence significative de durée des gestes entre la parole et le chant ($t = -1.7264$, $df = 234.615$, p-value $= 0.0856$) avec une durée moyenne de 1.11 sec dans la parole et de 1 sec dans le chant.

En ce qui concerne l'alignement des gestes avec la parole, dans la mesure où les actes de langage comme la proposition, par exemple, sont souvent répétés, il est difficile de déterminer quel groupe verbal constitue l'affilié lexical du geste. Mais si l'on suit les

travaux de Heath & Luff (2007), l'on peut supposer que l'apogée du geste qui, dans leur corpus, coïncide avec une syllabe nucléaire, coïncidera avec une syllabe longue dans notre annotation. C'est ce qui se produit pour 155 gestes sur 514. Dans ces cas, l'on peut considérer qu'il y a synchronie geste-parole, même si l'unité gestuelle est amorcée bien avant l'unité verbale qui contient la syllabe longue. Pour les autres gestes, il n'y a pas synchronie, mais anticipation ou retard du geste sur la parole. Afin d'en avoir une idée, nous avons assigné l'apogée de chaque geste à la syllabe longue la plus proche. Nous avons écarté les occurrences d'apogée situées à équidistance de deux syllabes longues (17 occ.), ainsi que celles situées à plus de 10 syllabes (53 occ.), ce qui correspond en moyenne au nombre de syllabes prononcés pendant la production d'un geste, mais ce chiffre pourrait être affiné. Les résultats montrent que pour 139 occurrences, le geste est produit en retard par rapport à la parole, alors qu'il est produit en anticipation dans 150 cas.

Le compte détaillé de chaque type de geste nous permet de dire qu'il n'y a aucun effet du type de geste sur le timing avec la syllabe longue la plus proche. Les pointages simples, les pointages complexes et les autres types de geste sont produits de manière proportionnelle en anticipation, en retard et en synchronie.

En ce qui concerne la répartition des types de geste par acte de langage, là encore, les types de geste sont également répartis entre les différents actes de langage et ne sont donc pas spécialisés. Par contre, le test de proportion montre qu'il y a 2 fois plus de gestes notés « autre » dans la répétition de la proposition que dans les autres actes de langage (X-squared = 7.5219, df = 1, p-value < 0.01). De fait, une proposition répétée est souvent accompagnée d'un simple battement dans la même direction que le pointage précédent.

On note en revanche quelques seuils de significativité entre le type d'acte de langage correspondant temporellement à l'apogée du geste et le timing de l'apogée par rapport à la syllabe longue la plus proche. On note que le geste qui accompagne l'acceptation du prix est plus souvent produit en anticipation par rapport à la syllabe longue la plus proche que les gestes produits sur les autres actes de langage (X-squared = 5.9042, df = 1, p-value = 0.01). On note également que les gestes qui accompagnent les actes d'encouragement et la simple mention de l'incrément ont une apogée le plus souvent produite en synchronisation avec une syllabe longue (actes d'encouragement : X-squared = 3.8526, df = 1, p-value < 0.05 ; mention de l'incrément : X-squared = 10.0446, df = 1, p-value < 0.01). Les gestes dont l'apogée coïncide avec les autres actes de langage ne montrent aucune régularité dans leur timing avec la syllabe longue la plus proche.

4 Conclusion

Dans cet article qui traite de la vente aux enchères chantée pratiquée aux Etats-Unis, et qui s'appuie sur un corpus vidéo de ventes pratiquées lors de concours, nous avons vu que la partie chantée de l'enchère diffère de la partie parlée sur le plan de l'intonation. La F0 est globalement plus élevée que dans la parole et la plage intonative des locuteurs est compressée entre 175 et 200 Hz pour les hommes, alors que la F0 comprend deux pics, l'un autour de 200 Hz, l'autre autour de 250 Hz chez les femmes. Cette

compression de la plage intonative est en grande partie due à la rapidité du débit des locuteurs, qui s'exprime dans la partie chantée de l'enchère, par une réduction phonémique (élision ou réduction de phonèmes, mais aussi réduction de la durée moyenne des phonèmes), ainsi que par une réduction de la durée moyenne des pauses, plus que par une réduction de leur nombre.

La gestualité nécessairement produite dans ce type d'interaction (gestes de pointage simple, de pointage complexe, ou autre type de geste) ne présente pas de différence par rapport à ceux de la parole en termes de durée ou d'amplitude. En revanche, la forte augmentation du débit de parole a un impact sur les gestes en termes d'alignement avec le verbal. Si l'apogée gestuelle concorde avec une syllabe longue pour un tiers des gestes produits dans le chant – ceci correspondant à ce que l'on rencontre dans les enchères traditionnelles – elle est produite en anticipation pour un autre tiers et en retard pour le dernier tiers. On observe cependant une certaine régularité selon le type d'acte linguistique : l'apogée gestuelle est majoritairement produite en anticipation d'une syllabe longue dans l'acte qui comporte l'acceptation du prix. Cette acceptation est le plus souvent le fruit d'une longue négociation du commissaire-priseur avec le public et l'anticipation de l'apogée montre que le commissaire-priseur a déjà connaissance de l'acceptation de l'offre au moment où il formule le prix verbalement une dernière fois, ce qui lui sert de tremplin pour passer à l'incrément suivant. L'apogée gestuelle est produite en synchronisation avec une syllabe longue dans les actes d'encouragement à accepter l'enchère, ainsi que dans la mention des incréments. Ces actes s'inscrivent en rupture par rapport au rythme de l'enchère – proposition, acceptation, nouvelle proposition, etc. – et il n'est donc pas étonnant que même si ces encouragements et ces mentions d'incrément sont chantés aussi, on observe une répartition différente des syllabes longues et brèves qui permet un meilleur alignement gestualité-parole.

Références

HEATH, C. et LUFF, P., (2007). Gesture and institutional interaction: Figuring bids in auctions of fine art and antiques. *Gesture* 7(2), pages 215-240.

HEATH, C. et LUFF, P. (2011). Gesture and Institutional Interaction. *In* (Streeck *et al.*, 2011), pages 276-288.

KUIPER, K. (1992). The Oral Tradition in Auction Speech. *American Speech* 67(3), pages 279-289.

KUIPER, K. (2000). On the Linguistic Properties of Formulaic Speech. *Oral Tradition* 15(2), pages 279-305.

KUIPER, K. et HAGGO, D. (1984). Livestock auctions, oral poetry, and ordinary language. *Language Society* 13, pages 205-234.

KUIPER, K. et TILLIS, F. (1985). The Chant of the Tobacco Auctioneer. *American Speech* 60(2), pages 141-149.

MCNEILL D. (2005). *Gesture and Thought*. Chicago and London : The University of Chicago Press.