PEGASUS - Communication Expressive : Analyse des données pour E-Clone

Aubergé V., Audibert N.*, Rilliard A.*, Rossato S.*, Tcherkassof A.*[+]
* Institut de la Communication Parlée (INPG – Université Stendhal / CNRS)
[+] Laboratoire de Psychologie Sociale (UPMF)

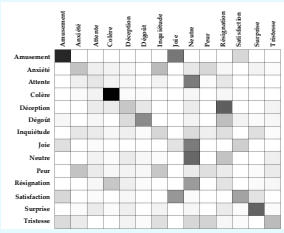
Résumé

Les stimuli actés produits par l'un des locuteurs du corpus « Sound Teacher » ont fait l'objet d'une évaluation perceptive par 26 sujets en conditions audio seul et audiovisuel, corroborant les résultats de la littérature, et montrant une différence inter-juge pour la reconnaissance de l'expression du dégoût en condition audio seul. De plus aucun effet significatif de la longueur des stimuli n'a été trouvé.

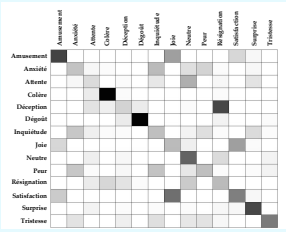
Nous avons également testé sur ce corpus la consistance de l'estimation du paramètre de qualité de voix NAQ, proposé par Campbell comme la 4ème dimension prosodique, ce qui laisse apparaître une dépendance phonémique de ce paramètre, ainsi qu'une surestimation systématique de NAQ sur le phonème [a].

Les contours des différents paramètres prosodiques ont été reproduits séparément par analyse / resynthèse sur des données actées bien reconnues lors de l'évaluation perceptive. Les stimuli ainsi resynthétisés ont été évalués par 16 sujets, d'où il semble ressortir que les contours de F0 donnent plus d'information sur les émotions positives, et les contours de qualité de voix et de durée (qui n'ont pu être séparés avec la méthodologie utilisée) sur les émotions négatives.

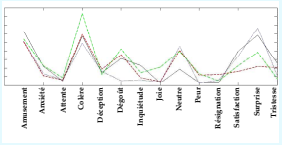
Evaluation perceptive



Audio seul



Audiovisuel



Pas d'effet significatif de la longueur des stimuli

Quels paramètres? Quelle morphologie?

quantités statiques multi-dimensionnelles
vs.
contours gradients multi-dimensionnels (« Gestalts »)

Même si la prosodie a une morphologie en contours, des mesures quantitatives et des corrélations entre ces valeurs peuvent être **suffisantes pour une tâche de reconnaissance.**

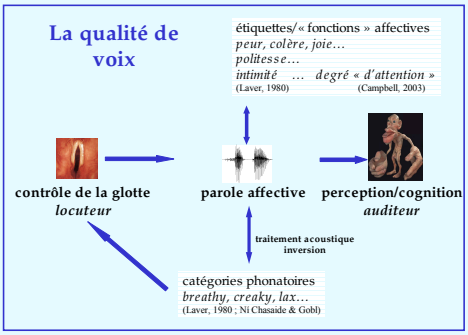
(cf. approche tonologique pour décrire les contours de la prosodie linguistique)

mais pas en génération

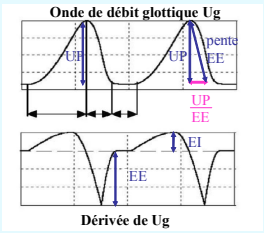
si les contours sont porteurs d'information, leur reconstruction à partir de formes statiques est impossible

⇒ **besoin de contours**

La qualité de voix



Calcul du paramètre NAQ



Estimation de NAQ :
phase de fermeture de la glotte normalisée par F0

$$NAQ = \frac{UP}{EE} \times F0$$

(Alku et al., 2002)

Estimation de l'onde de débit glottique par filtrage inverse

(Mokhtari et al., 2002)

NAQ comme une dimension de la prosodie des affects

le Quotient d'Amplitude Normalisé (NAQ) a été proposé comme caractéristique:

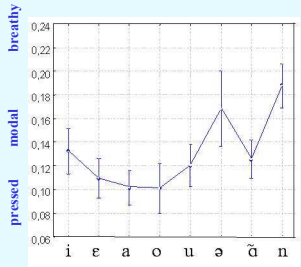
- **de l'articulation *breathy*** par Alku et al. (2002)
- **du degré d'attention porté à la voix** par Campbell (2003)

NAQ est calculé à partir de mesures statiques sur un très large corpus, il varie en fonction des situations d'élocution (interlocuteur)

Problématique de la génération : faire une analyse morphologique sur une unité de parole, obtenir des contours du paramètre NAQ

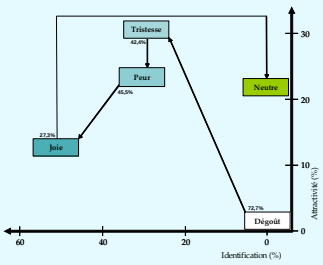
Questions :
- Y a t'il une influence du phonème sur NAQ ?
- Existe-t-il des artefacts dus au filtrage inverse dans le calcul de NAQ ?

Influence du phonème sur NAQ

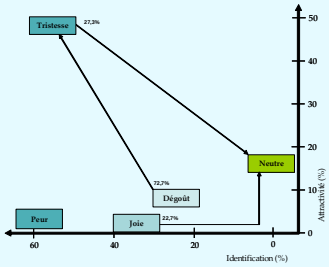


- **voix *pressed* (0.08-0.11), *modal* (0.11-0.17) ou *breathy* (0.23-0.35)** (Alku et al., 2002)
- **effet significatif du phonème sur NAQ**
- **même proportion de [ə] ajoutés mais avec un NAQ élevé pour le locuteur 1 uniquement**
- **NAQ élevé pour les nasales: la nasalité renforce les basses fréquences, d'où une pente spectrale élevée**

Evaluation séparée des paramètres par analyse / resynthèse



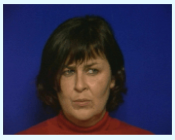
fréquence fondamentale



qualité de voix + durée

Timing des gestes faciaux vs. vocaux

Communication non verbale

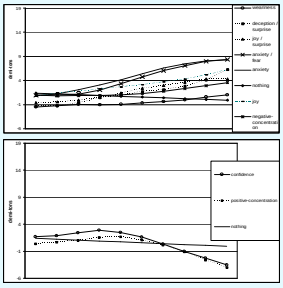


Préparation à l'action
Expression des états mentaux
des affects et cognets
=> Feeling of knowing

Communication verbale

Un seul contrôle bimodal
= 1 expression bi-modale cohérente
ou
Un contrôle spécifique pour chaque modalité
= 2 expressions harmonieuses

Prosodie émotionnelle: gradiente ou par contours?



Bänziger et al. (2003) ont obtenu des résultats perceptifs montrant que seule la valeur moyenne de F0 est perceptivement significative et que la hauteur générale de F0 peut être reliée à des valeurs distinctes d'activation et d'arousal. Ces deux caractéristiques sont certainement les plus pertinentes, mais nous suggérons que des formes extralinguistiques plus fines peuvent être masquées par les contours linguistiques porteurs qui, quoique supposés identiques pour les énoncés longs choisis, les écraseraient. Notre objectif est de répliquer une expérience perceptive similaire sur des unités linguistiques minimales et pragmatiquement gelées. Nous présentons ici des contours de F0 normalisés et stylisés sur 10 points. Lorsqu'on compare aux valeurs de référence (pour la parole actée aussi bien qu'authentique), différents types de contours peuvent être observés, suffisamment simples pour pouvoir expérimenter divers types de stylisation dans des expériences perceptives à venir. Ces tests pourront déterminer dans quelle mesure les contours (notamment au niveau des glissandos) sont sensibles aux différentes étiquettes émotionnelles. Bien qu'il nous soit à l'heure actuelle impossible en l'absence de validation perceptive des stimuli spontanés de conclure à la pertinence des contours pour décrire la parole émotionnelle, on peut noter qu'on observe, pour la parole actée et authentique, un contour sans valeur émotionnelle (noté *nothing*) exactement comme prédit sur des unités minimales linguistiques gelées, et que les contours semblent bien varier en fonctions des variations d'état émotionnel.