Early meaning before the phonemes concatenation? Prosodic cues for Feeling of Thinking

Anne Vanpé [1], *Véronique Aubergé* [1, 2]

[1] GIPSA-Lab, CNRS, Grenoble, France
[2] LIG, CNRS, Grenoble, France

1. Problematics and literacy overview

The non lexical sounds that are non phonological but prosodically relevant, are produced both during or outside the talk turn. They have been observed both in listener feedbacks in backchannel and in the feedbacks of the speaker, implied in a human/human or human/machine interaction (Schröder et al, 2006). The "mouth noises", interjections, fillers, grunts, bursts etc. have been studied for their emotional functions (Scherer, 1994; Schröder, 2003; Campbell, 2004) or for their pragmatic functions in dialog (Ameka, 1992; Wichmann, 2002: Ward, 2006; Poggi, 2008). They can express emotions, intentions, attitudes and cognitive/mental states and processing (like concentration, hesitation about an answer, etc.) that we name Feeling of Thinking - FoT.

This paper presents a subtle annotation, description and organization of such non lexical sounds from a large spontaneous corpus: 6 French speakers (3 males, 3 females, 3h45 video-taped), from the Sound Teacher/E-Wiz HMI corpus based on language learning tasks, emotionally induced (Aubergé et al, 2003).

The questions are: why, how and when the acoustic modality is used, outside the speech production? Can the non-lexical sounds occurrences directly be related to FoT or to the task organization? As for the acoustic nature of these sounds, it can supposedly be described following the degree of complexity of the prosodic control: (1) for "bio-physiological" sounds: no control, (2) for non phonetic sounds: duration and illocutory force, then supra-glottal voice quality, then F0 and glottal voice quality (3) for phonetic or phonological sounds (like interjections): all the dimensions of prosody. A preliminary experiment (Signorello et al, 2010) has shown that French vs. Italian listeners can decide which language corresponds to the non-lexical sound as soon as a minimal prosodic control appears (that is the duration and illocutory force -kind of energy- of non phonetic sounds). Therefore, the present work raises the question where does the language code begin: with the double articulation, or with "pure prosodic word"? Is language code originally based on sound symbolism?

2. Labeling methodology

Labeling the forms of expressions, as well as annotating the values of expressions, are crucial issues. In order to get a complete and subtle labeling, free from any theoretical filter, the labeling has been performed without any knowledge about self-annotations and induction context. Thus, a bottom-up approach was adopted, avoiding any a priori knowledge about the nature or the function of the micro-events A distinction has been made between interjections (pre-lexical items built with language phonemes), and other voice events (*e.g.* bursts, breathy noises, clicks, clearing one's throat, moans, etc.).

Interjections have been classified according to phonetic features: vocalic phoneme (labeled "V", *e.g.* [ø:] "euh"), consonantal phoneme (labeled "C", *e.g.* [m:] "mmh") and combinations of vocalic and consonantal phonemes ("CV", e.g. [bø:] "beuh", and "VC", *e.g.* [ø:m:] "euh mmh"). When an interjection was composed of more than one consonantal or vocalic phoneme, it was labeled "Comb" (for "combination", e.g. [bʃE)m:] "ben mmh", [ula] "ouh

là"). Complex voice events were labeled as well as simpler ones. Voice quality has also been labeled if necessary.

The other than interjection voice events, globally named "mouth noises" are described according to their articulatory/acoustic nature, including possible voice/sound quality. They are also classified into (1) produced by the subjects during an ingressive or an egressive airflow; (2) produced with a continuous airflow (*e.g.* strong inspiration, moans), restricted airflow (friction), or with at least one airflow block (including glottal stop, *e.g.* clicks, bilabial plosives). We used the second parameter because it involves a tenseness followed by a release of the subject. We found also two other kinds of mouth noises, linked either to an interaction between tongue and lips or to "swallows".

3. Occurrences of non-lexical sounds analysis

As expected, the most frequent interjections are vocalic (316 occurrences, 64.5% of the interjections), and mainly the vocalic filler [ø:] ("euh"). It is the case for all subjects, and including or not "V-variants" (carrying voice quality). 11.8% of the interjections are "VC" interjections, 10.6% are "CV", 6.9% are "C", and 6.1% are more complex combinations.

The inventory of mouth noises (Table 1) shows that 76.1% of mouth noises are shared between only 5 out of the 13 defined types. Notice that it concerns mouth noises whose production requires a quite simple articulatory control.

Otherwise, the mouth noises frequency seems to depend on their nature (Table 2): the most frequent mouth noises are produced in an ingressive and blocked airflow (37%). Globally, 59% of mouth noises are produced in an ingressive airflow. Among them, 63% are "blocked", while most of egressive mouth noises (59%) are produced in a continuous airflow. In parallel, 78% of "blocked" mouth noises are produced in an ingressive airflow, while "continuous" mouth noises are well-distributed in terms of ingressive / egressive airflow. Otherwise, only 5% of the mouth noises are produced with a restricted airflow.

More precisely, the analysis of mouth noises in terms of perception of their voicing and voice/sound quality (Table 3) shows that only 5% of mouth noises is voiced. Most of them (88%) don't carry any voice quality. 6 out of the 15 occurrences that carry one are whispered, and 5 are sighed. Notice also that 73% of the voiced mouth noises occurrences that carry a non modal voice quality are produced in a continuous airflow. Moreover, it is possible to find some "sound quality" carried by non voiced mouth noises (35 occurrences). 71% of them are sighed. As for voiced mouth noises, these occurrences are mainly produced in a continuous airflow (86%). However, their proportion to the whole non voiced mouth noises is low (only 2%).

Analysis shows (Table 4) that the production of voice events is continuous but not frequent (from 8.3 to 10.7 per minute) during "stimulus-responses" tasks, and continuous and frequent (22.5 per minute in average) during "sub-results" tasks. Interestingly, their productions are not continuous during phase results (when the induction is the strongest), and their production frequency seems to be dependent on the kind of induction: it is especially high during the very negative induction.

4. References

Ameka, F. (1992). "Interjections: The universal yet neglected part of speech", Journal of Pragmatics, 18, 101-118, 1992.

Aubergé V., Audibert N., Rilliard A., (2003). "Why and how to control the authentic emotional speech corpora?", Proceedings of Eurospeech, Genève, 321-325.

Campbell, N. (2004). "Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language", Languages Resources and Evaluation, 39, 109-118

Poggi, I. (2008). “The language of interjections”, Multimodal Signals: Cognitive and Algorithmic Issues, COST 2102 School, Vietri, Italy, 170-186,

Scherer, K.R. (1994). “Affect bursts”. In S.H.M. van Goozen, N. E. van de Poll & J.A. Sergeant (Eds.), Emotions, Hillsdale (NJ, USA), Lawrence Erlbaum, 161-193,.

Schröder, M. (2003). “Experimental study of affect bursts”, Speech Communication, 40(1-2), 99-116.

Schröder M., Heylen D., Poggi I. (2006). "Perception of non-verbal emotional listener feedback", *Proc of Speech Prosody 2006*

Signorello, R., Aubergé, V., Vanpé, A., Grandjon, L., Audibert, N. (2010). “A la recherche d’indices de culture et/ou de langue dans les micro-événements audio-visuels de l’interaction face à face”, Proceedings of WACA 2010, Lille, France, 69-76.

Ward, N. (2006). “Non-lexical conversational sounds in American English”, Pragmatics & Cognition, 14(1), 129-182 (54).

5. Tables

Table 1. *Inventory of mouth noise types, ordered by frequency*

Mouth noise type	Number of occurrences	Percentage of total
inspiration	577	22,9%
articulator slackening	477	18,9%
expiration	393	15,6%
ingressive occlusion_inspiration	251	10,0%
swallowing	220	8,7%
plosion	151	6,0%
respiration releasing	127	5,0%
brutal expiration	126	5,0%
tongue-lips interaction	92	3,6%
tongue click	76	3,0%
friction	13	0,5%
clearing one's throat	10	0,4%
moan	8	0,3%
Total	**2521**	100,0%

Table 2. *Number and percentage of mouth noises according to their airflow type (ingressive vs. egressive, and “blocked”/“restricted”/“continuous”).*

Airflow type	Egressive airflow			Ingressive airflow			swallows	tongue-lips interaction	Total
	blocked	restricted	continuous	blocked	restricted	continuous			
Number of occurrences	239	69	437	861	52	461	167	72	**2358**
Percentage of the total	10%	3%	19%	37%	2%	20%	7%	3%	100%
Sub-total	745			1374					

Table 3. *Number of mouth noise occurrences according to their voicing and voice/sound quality cues, and depending on their airflow type.*

Mouth noise airflow type -->	voice/sound quality	block	restricted	continuous	Total
Non voiced	modal	1300	84	817	**2200**
	whipered	1		1	**2**
	creaky			4	**4**
	sighed	3	1	21	**25**
	shakily			4	**4**
Non voiced total		1304	85	847	**2236**
Voiced	modal	26	32	49	**107**
	whipered	2	1	3	**6**
	creaky		1	2	**3**
	murmured			1	**1**
	sighed			5	**5**
Voiced total		28	34	60	**122**
Total		**1332**	**119**	**907**	**2358**

Table 4. *Average number of voice events per minute, according to the recurrent task during which they are produced, and the different phases of the scenario.*

Recurrent task	Phase 1: positive induction	Phase 2: very positive induction	Phase 3: negative induction	Phase 4: very negative induction	Total
sequence (heard stimulus / answer)	8,3	10,8	9,7	10,7	**9,4**
reading of sub-results + comments	20,8	25,7	24,6	22,8	**22,5**
reading of phase results + comments	16,9	10,7	2,4	21,1	**20,7**
Total	**10,7**	**15,2**	**13,6**	**14,5**	**12,6**