Vocal effort in situation

Jean-Sylvain Liénard

Limsi-Cnrs, Orsay, France

Pevoc9, Marseille 02-09-2011

- Section 1: Introduction, historical background
- Section 2: Vocal effort and communication situation
- Section 3: Acoustic features of the vocal effort
- Section 4: Vocal effort and sound technologies
- Section 5: Conclusion

1. Introduction

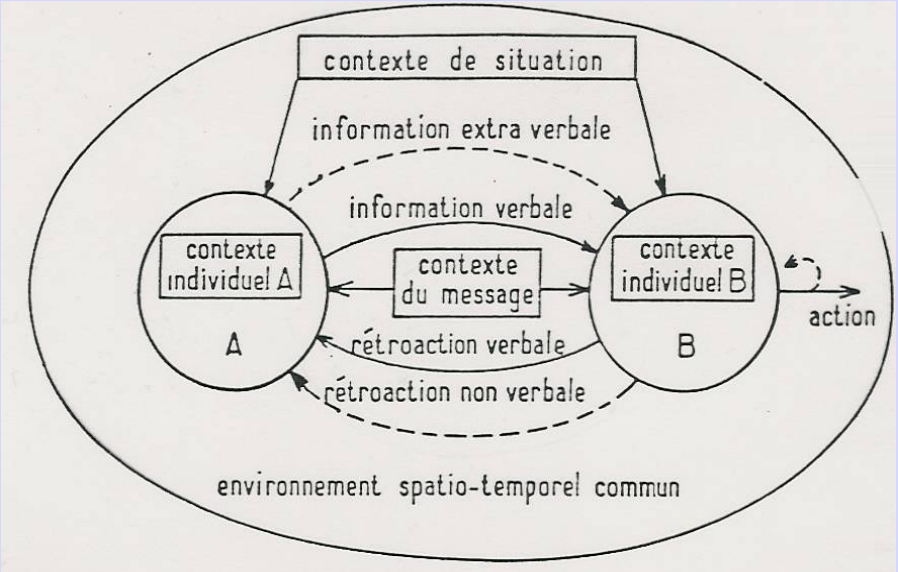
1.1 Historical background

- In the 70's: towards engineering speech communication
  - Phonemes as pearls on a necklace, syntax, limitations due to "speech variability"
- Observation 1: speech IS NOT an oralized version of writing
  - writing is an abstract, formalized notation of speech; non-linguistic info is lost
- Observation 2: speech perceptible variations carry useful information
  - Instead of looking for invariants, let us look for the causes and significance of the variations
- More recently (90's): why and how do people communicate ?
  - To control and synchronize their behaviours or minds
  - They take advantage of the common context and situation
  - They anticipate at any time scale and pre-select the information they need

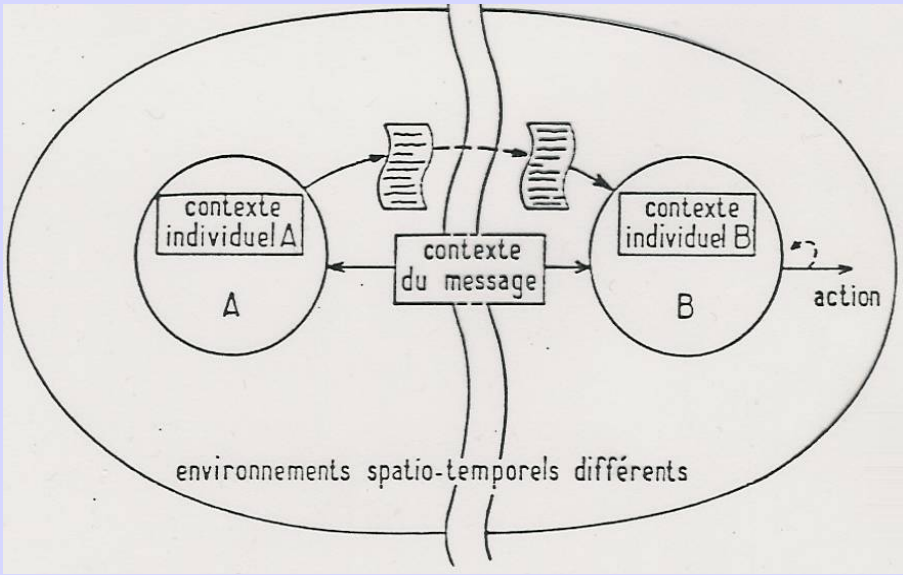
→ situated perception/cognition

1.2 Speech vs writing [1]

Oral communication



Written communication



→ Speech communication is a situated, interactive activity

[1] From Liénard, Les processus de la communication parlée, Masson, Paris, 1977

1.3 Bjorn Lindblom H&H theory

- Formant reduction in some rapid vocalic transitions
  - correct vowel may be perceived although target formant not attained

- Hypo & Hyper articulation (1)
  - Speaker subjected to 2 contradictory constraints
    - Speaker has to ensure intelligibility for the listener → tendency to hyperarticulation, but this has a cost in terms of "articulatory energy"
    - As any human, speaker has a natural tendency to economy → hypoarticulation
  - Degree of articulation may evolve in time along the hypo-hyper continuum

(1) Lindblom 1990: Explaining phonetic variation: a sketch of the H&H theory. In Speech Production and Speech Modelling, eds Hardcastle and Marchal, Kluwer, 403-439.

2. Vocal effort and communication situation

- Natural situations
  - Situations where no sound device is used (microphone, loudspeaker, telephone, computer…)
  - See in section 4 the new situations due to sound technologies

2.1 Speaker adjusts his VE in order to reach a given listener

- Speaker knows what listener/interlocutor he intends to address
  - Sees him/her (often)
  - Or, in the presence of several potential addressees, has selected one of them
  - Knows of his/her possible auditory difficulties
  - Knows of the acoustic transmission conditions (distance, reverberation, noise)

→ Speaker immediately chooses the right VE to use in the given situation
  - Uses pre-stored (learned ?) knowledge of a similar situation to tune the vocal apparatus

2.2 Interlocutors know what VE to expect from the speaker

- It all depends on the situation:

  Example 1: One-to-one conversation
  - "normal" voice for each: just loud enough to get heard by interlocutor
  - any deviation from this "normal voice" (too loud, whispered, fading, harsh…) means something

  Example 2: professor delivering lecture to 20-30 students
  - Everyone expects the prof's voice level to be high enough to reach the farthest students
  - No one feels addressed in particular. Here "normal voice" is rather loud

  Example 3: several persons silently working in a room; newcomer emits some oral request
  - If newcomer's voice lower than norm: closest people know they are the ones addressed

→ Vocal behaviour - and Vocal Effort - is adopted and interpreted by all interlocutors according to the situation

2.3 Voice ranges and voice types

Voice range:

Conditions to be specified: free (-6)/open field (-4), closed room, noise level and structure

Voice dynamics: from a few cm to some 200 m: ratio $\sim 10^4$ → 50-60 dB or more

Is that the range of variation due to the vocal effort ? Yes and no: 3 distinct voice types

Voice types (modes ?):

Whispered (low, high): a few cm to 1 m; intimate use. Proximity voice

Voiced (weak, usual, loud): 0.5 to 10 m. Conversational voice

Shouted: (mid-range, long-range): 10 to 200 m: exceptional; screams, emergency calls, high-level noise: restricted speech communication capability

Level variations in conversational voice (from weak to loud) are contained within a 15-20 dB range

(Not to be confused with S/N ratio)

3. Acoustic features of the vocal effort

Several studies since the 70's

Some authors: Rostolland[(1)], Schulman [(2)], Junqua[(3)], Liénard[(4)], Traunmuller[(5)], Garnier[(6)] , Zahoric[(7)] (both texts [(6)] and [(7)] give exhaustive lists of references).

Studies differ in

Purpose:

- Intelligibility of shouted voice in working environments (Rostolland)
- Articulatory correlates of loud speech (Schulman)
- Lombard effect (Junqua)
- Spectral properties of speech at different ranges (Traunmuller, Liénard & Di Benedetto)
- Vocal straining (Garnier)
- Auditory distance estimation, for any sound source incl speech (Zahoric)

Vocal effort elicitation:

- Asking subjects to produce loud or shouted voice
- Immersing subjects in some controlled noise
- Placing subject at increasing distance of a given listener/interlocutor

(1) Rostolland 1982: Acoustic features of shouted voice, Acustica **50**, 118-125
(2) Schulman 1989: Articulatory dynamics of loud and normal speech. JASA, **85**, 295-312
(3) Junqua 1996:The influence of acoustics on speech production: (…) the Lombard reflex, SpeechCom **20**, 13-22
(4) Liénard and Di Benedetto 1999: Effect of vocal effort on spectral properties of vowels, JASA **106**, 411-422
(5) Traunmuller and Eriksson 2000: Acoustic effects of variation in vocal effort by men, women and children, JASA **107**, 3438-3451
(6) Garnier 2007: Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal, Thèse de doctorat, Univ. Paris 6
(7) Zahoric et al 2005 Auditory distance perception in humans: a summary of past and present research, Acta Acustica **91**, 409-420

3.1 Overview of the Liénard-Di Benedetto study

Objectives

- To study variability due to the VE in usual communication conditions: naive subjects, furnished room, no repetitions, no artificial constraints, no noise
- Simple and significant data: 9 French isolated vowels, 10 speakers, 2 main sources of variability (gender, vocal effort)

Experimental setup

- Speaker seated, omnidirectional microphone located 30 cm from mouth, locked input level
- Operator located successively at 1.5m ("Normal" condition), 0.4m ("Close" condition), 6m ("Far" condition.

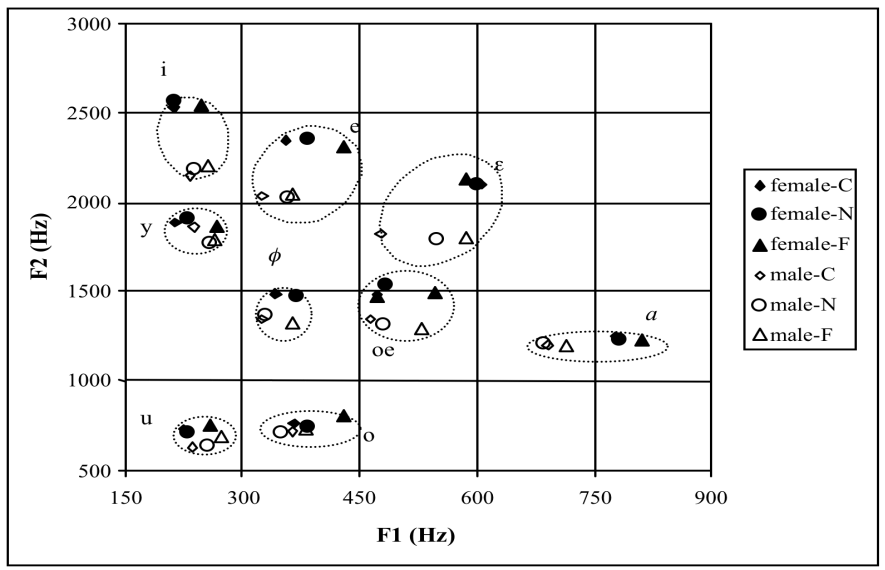
Perceptual validation

- Evaluation by 5 listeners (separate group)
- No repetition allowed, listening level fixed at comfortable by subject, then locked
- 3 answers requested: perceived vowel, perceived VE, perceived gender, "don't know" allowed
- results
  - 9.3% error on vowel label
  - 7% error on gender
  - No effect of the distance condition
  - 41% error on VE (distance): better than random: 66% (other studies have shown that VE perception was a good estimator of distance)

3.2 Results: effects of VE

- Token intensity correlated with distance condition; increase of 2.25 dB when distance doubles (closed room)
- F0 and F1 vary significantly as a function of the distance condition
  - F0 increases with VE: 5.1 Hz/dB (high correlation)
  - F1 increases with VE: 3.5 Hz/dB (moderate correlation)
- No significant effect on higher formants

- Formant intensities augment faster than token level : spectrum mid and high parts get reinforced with VE



3.3 Remarks on VE-related acoustic features

All studies agree on increase of: intensity, spectral emphasis, F0 and F1

Spectral balance

Vowels: Spectral emphasis + F0 and F1 upwards shift → strong variation in spectral balance in Barks

Consonants: bursts and frication noises unchanged, perceptively wiped out at long distance

Is there an influence of VE on prosody ?

Open question, to investigate with 2 choices in mind:

1. What is the situation (speaker and listener) ?
   - Lombard and distant speech situations deeply different
2. What is the type of voice used by the interlocutors ?
   - Whispered and shouted voice induce specific situational constraints
   - Interlocutors must have sufficient "intonational degrees of freedom" to incorporate the linguistic and non-linguistic info

4. Vocal effort and sound technologies

4.1 New communication situations

- Telephone
  - Interlocutors immersed in separate contexts, communicate anyway
  - Possible conflict between local and distant situations

- Public address
  - Speaker physically far from listeners but his voice looks close
    → VE chosen so as to suggest proximity without getting too close

- TV or broadcast presentation
  - Presenter chooses VE so as to suggest some presence to the targeted average listener, without looking intrusive
  - Target listener has expectations about how close he/she wishes the speaker to be located
  - Changing the "volume" control does not change the perceived distance

→ New: auditory distance no longer identical to physical distance

→ Natural situations remain; the new ones simply add up to them.

4.2 Human-machine communication

Speech synthesis from the text

- Intelligible, unlimited vocabulary, several (pre-stored) voices available
- is just an oralization of the text: "dead speech". Quicky boring.
- How to define a situation interactively linking machine and listener ?
- Some research efforts to provide interaction capabilities[1]

Speech recognition

- Works well in extremely limited applications, robustness problems
- Research not aware of factors such as prosody and situation
- However, investigating variability due to VE may help to improve performance

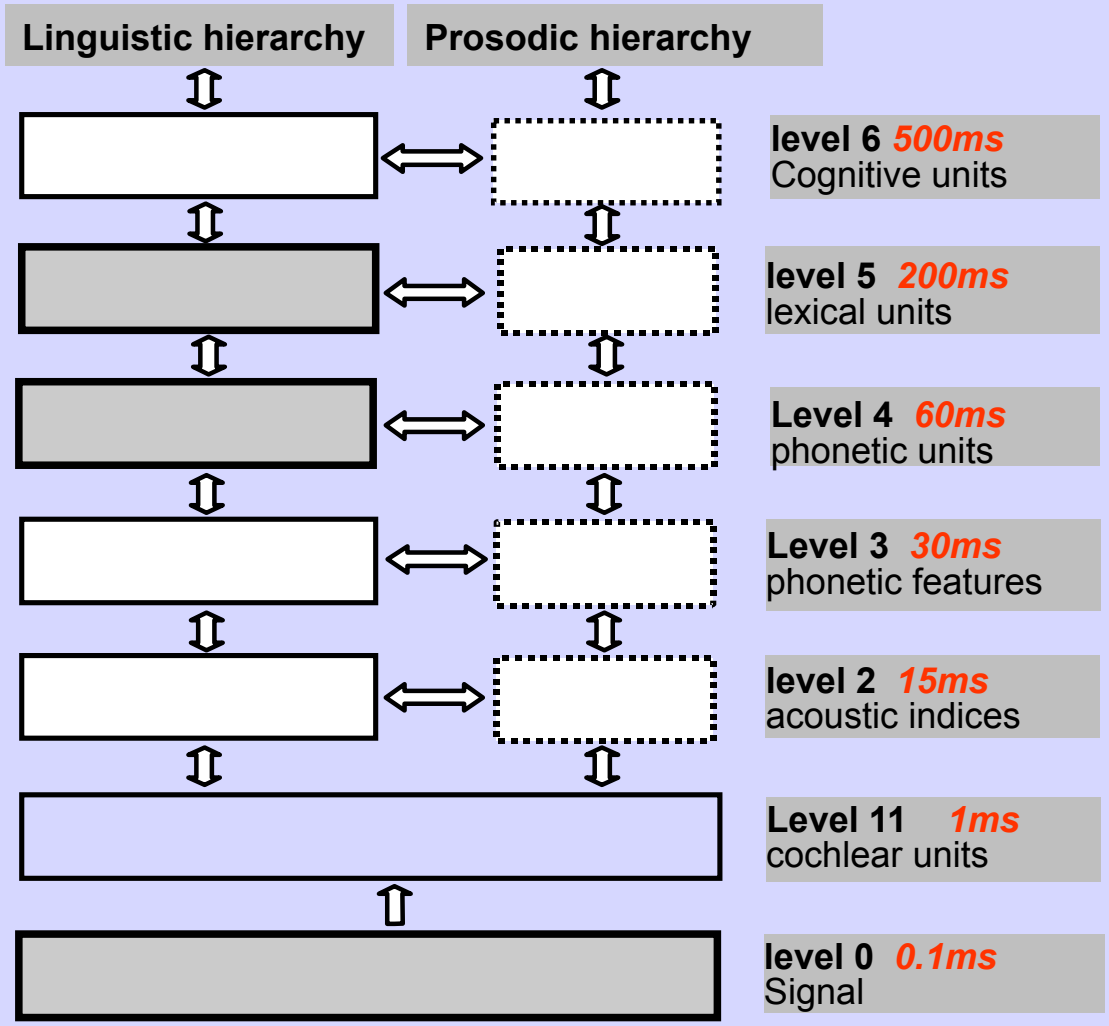
[1] Moore, R. K., & Nicolao, M. (2011). Reactive speech synthesis: actively managing phonetic contrast along an H&H continuum, *17th ICPhS,* Hong Kong.

5. Conclusion

Vocal effort

- Is one of the main sources of variability in speech communication, even within a reduced variation range
- Is determined by physical and psychological situation of interlocutors
- Is fully automated and unconscious, at least in the conversational range
- Has been widely underestimated until now in acoustic/phonetic studies

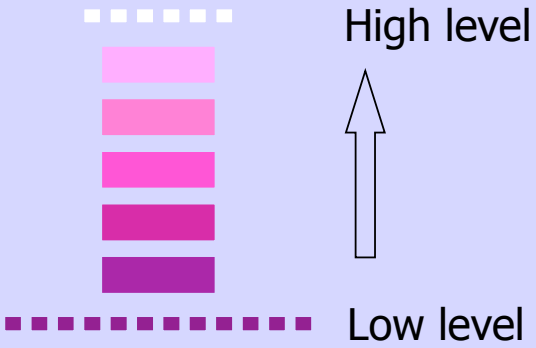
1.4 Speech and voice perception



- *abstraction levels are tied to time resolution*
- *joint processing of all types of info*
- *at each level the description of the perceptive content is "complete" (linguistic and non-linguistic)*
- *descriptors more and more independent as level increases*
- *two information flows coexist: bottom-up and top-down*

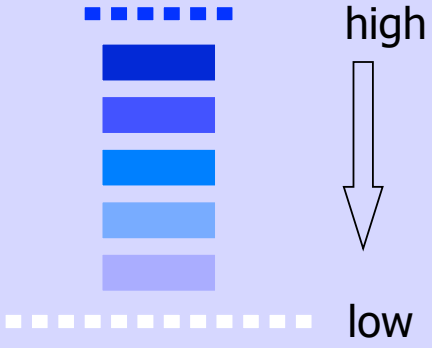
1.5 Functioning modes

Bottom-up



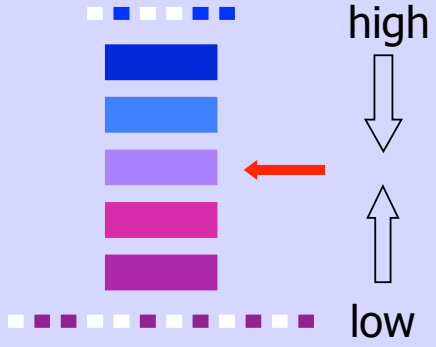
- low-level info dominates the process
- no anticipation
- streaming, pop-up, intrinsic descriptors, Gestalt

Top-down



- High-level info dominates the process
- full anticipation
- attention and knowledge governed by upper level

Double flow



- low-level info is partial, as well as high-level info
- an intermediate level dominates the process
- possible conflict

1.6 A cognitive model of individual perception and action