‘Seeing tunes.’ The role of visual gestures in tune interpretation
Introduction (1/3)

- Strong influence of visual cues upon speech perception in various aspects of communication typically associated with verbal prosody
  - Boosting the perception of prominence and focus (Cavé et al., 1996; Hadar et al., 1983; Krahmer & Swerts, 2007; Swerts & Krahmer, 2008; Dohen & Løevenbruck, 2009)
  - Expressing affective meanings such as uncertainty (Krahmer & Swerts, 2005) and frustration (Barkhuysen, Krahmer & Swerts, 2005), in face-to-face grounding (Nakano et al., 2003) and question intonation (Srinivasan & Massaro, 2003)

- Many studies have described a correlated mode of processing, whereby vision partially duplicates acoustic information and helps in the decoding process
  - Visual information provides a powerful assist in decoding speech in noisy environments, particularly for the hearing impaired (Sumby & Pollack, 1954; Breeuer & Plomp, 1984; Massaro, 1987; Summerfield, 1992; Grant & Walden, 1996; Grant, Walden & Seitz, 1998; Assmann & Summerfield, 2004)
Most studies have found a weak visual effect relative to a robustly strong auditory effect
- Prominence and focus marking (Scarborough et al., 2009; Krahmer & Swerts, 2004; Krahmer et al., 2002; Krahmer & Swerts, 2004; Dohen & Lœvenbruck, 2009)
- Question marking (House, 2002; Srinivasan & Massaro, 2003)

Few studies have found that vision is more important
- Rating the attitude of a speaker (like, neutrality, dislike) (Mehrabian & Ferris, 1967)
- Estimating the uncertainty of a speaker (Swerts & Krahmer, 2005; Dijkstra, Krahmer & Swerts, 2006)

So it is still unclear how crucial visual cues are compared to auditory cues in expressing intentionality
How important are facial gestural correlates and pitch accent cues in the listeners’ decisions between statements and questions?

In Catalan, a pitch range difference in a rising-falling nuclear configuration is the main intonational cue for the pragmatic distinction between contrastive focus statements and counter-expectational questions (Borràs-Comes, Vanrell & Prieto, 2010). Alignment LH properties are the same (Prieto 2002).

We will use congruent and incongruent AV combinations which will allow us to assess the relative perceptual importance of visual vs. auditory cues.
No previous research dealing specifically with the facial gestures that characterize counter-expectational questions and focus meanings in Catalan.

In order to decide which gestural patterns would be used as target facial expressions in our visual materials, 10 native speakers of Catalan (ages 20-47) were videotaped pronouncing both possible interpretations of the utterance.

Subjects were asked to read in an expressive way these two dialogues:

a. — Volies una cullera gran, no? — Petita, [la vull, i no gran].
   You wanted a big spoon, didn’t you? [I want a] little [one, not a big one].

b. — Jo la vull petita, la cullera. — Petita? [N’estàs segur?] I want a little spoon.
   [A] little [one]? [Are you sure?]
Production experiment (2/2)

- **Contrastive focus statement**
  - upward eyebrow movement
  - forward head movement

- **Counter-expectational question**
  - furrowing of the brows
  - squinting of the eyes
  - often accompanied by a head shake
2 identification experiments

- **Goal**: to test the relative importance of facial cues with respect to auditory cues

- **Procedure**: Subjects were presented with either congruent or incongruent AV target stimuli
  - Experiment 1
    - acoustic continuum of pitch range coupled with 2 video clips of a person’s face with his expression conveying each target meaning
  - Experiment 2
    - same stimuli, but this time with a visual continuum created through a digital image-morphing technique

- **General instructions**: “identify the intended meaning (statement or question) for each combined AV stimulus”

- 20 native speakers of Central Catalan participated

- Experiment set up by means of E-Prime 2. RTs also recorded
Exp. 1: visual materials

- A short utterance contain the target intonational and facial cues:
  *Petita* [pəˈti.tə] (‘small’-fem.)
  - Importantly, auditory and gestural peaks are both aligned with the accented syllable, that is, we can claim that they co-occur in time
- To generate the visual stimuli for the experiment, a native speaker of Catalan was videotaped several times. He imitated the 2 gestural patterns selected from the preliminary recordings
  - 2 exemplars were selected (with similar syllabic durations)
Exp. 1: acoustic materials

- Target sentences pronounced with a rising-falling intonational contour (L+H* L%) but different in pitch range

- One of the two auditory recordings was manipulated (Praat) in order to create a synthesized continuum of 11 steps (distance between each one = 0.6 semitones)
Experiment 1: results

- 2,200 responses (11 auditory steps × 2 visual sequences × 5 blocks × 20 listeners)

- Clear division depending on the visual material used

- Clearer responses for congruent AV combinations

- Visual effect
  - F (1, 2007) = 1306.798, p < .001

- Auditory effect
  - F (1, 2007) = 1.059, p < .001

- No AV interaction (p = .391)
Experiment 1: RTs

- Clear AV interaction
  - Lower RTs for congruent AV combinations

- AV congruency effect
  - $t_{(183)} = -3.619, p < .001$

- Visual effect
  - $F(1, 2173) = 6.362, p = .012$
  - No auditory effect ($p = .752$)

- AV interaction
  - $F(10, 2173) = 2.815, p = .002$
4 static images were extracted from the two visual materials of our Experiment 1
- one for the initial neutral gesture
- one at the beginning of the stressed syllable
- one at the beginning of the post-tonic syllable
- one at the end of the utterance

Then, a face morphing technique was applied to the 2nd, 3rd and 4th stills selected, in order to create 4 intermediate videos in between the 2 natural visual gestures
- With this technique, we can morph one face into another by marking key points on the first face, such as the location of an eye, and mark where these same points are located on the second face. The program then creates intermediate frames between the original faces (inbetweens)

Once we had the 6 sets of four images, we (re-)synchronized them with the audio
Experiment 2: results

- 3,600 responses (6 auditory steps × 6 visual sequences × 5 blocks × 20 listeners)

- Clear division depending on the visual material used
  - Enhanced effect of auditory stimulation when visual materials are more ambiguous

- Clearer responses for congruent AV combinations

- Visual effect
  - $F(5, 3404) = 289.617, p < .001$

- Auditory effect
  - $F(5, 3404) = 149.821, p < .001$

- No AV interaction ($p = .093$)
Experiment 2: RTs

- **Clear AV interaction**
  - Lower RTs for congruent AV combinations

- **Longer RTs when visual materials were more ambiguous**

- **AV congruency effect**
  - $t_{(180)} = -2.194, p = .029$

- **Visual effect**
  - $F(5, 3564) = 11.608, p = .012$

- **No auditory effect** ($p = .601$)

- **AV interaction**
  - $F(25, 3564) = 1.579, p = .034$
In sum... (1/2)

- Effect of both visual and auditory cues
  - In both experiments, listeners’ decisions mainly depend on whether the video component of the AV material shows facial expressions corresponding to a statement or a question
  - Thus the present results show that focus statements and questions can be discriminated predominantly from visual information, with auditory information (on the basis of an F0 pitch range contrast) playing a secondary reinforcing role
Effect of bimodal auditory+visual congruity

That is, when Catalan listeners rated a question-based visual stimulus occurring with an incongruent low-pitched auditory stimulus, an important latency appeared in the response, and vice-versa.

Enhanced importance of auditory stimuli when visual input is ambiguous

- Logistic regression: shallowest slopes for ambiguous visual stimuli
Conclusions

- In both experiments, the response frequencies given by Catalan listeners revealed a clear preference for visual cues when deciding between contrastive focus statements and counter-expectational questions
  - Gesture acts as an integral part of language comprehension and, as such, provides insight into fundamental aspects of prosodic interpretation

- There is an effect of bimodal audio + visual congruity
  - The strong effects of congruity both in patterns of results and in RTs represent a clear argument in favor of the view that gesture together with speech form a single integrated system
  - Also, there is an enhanced importance of acoustic stimulation when visual input is more ambiguous
An ongoing experiment using the gating paradigm has tested the perception of a set of gated utterances (broad and contrastive focus statements and counter-expectational questions) occurring in the 3 possible modalities: AV, AO and VO.

- VO and AV conditions trigger faster processing.
  - No differences appear depending on the presence of simultaneous auditory input.
  - Responses to the AO condition are quite late.
Further research

- Which is the relevance of target facial cues and their contributions to the statement-question judgments?

- Usage of a computer-generated 3-D talking head
  - Each intended gesture (eyebrow position, eyelid closure, and head movement) is manipulated separately
  - They appear on 4 levels of strength

- Pilot results
  - Very significant effect of eyebrow lowering (F (3, 5014) = 761.960, p < .001)
  - Interaction eyebrow*eyelid
We are grateful to Carme de-la-Mota, Marion Dohen, Eva Estebas-Vilaplana, Itziar Laka, Lluís Payrató, Josep Quer, Núria Sebastian-Galles and Marc Swerts for their comments. We would also like to thank Meghan Armstrong, Verónica Crespo-Sendra and Maria del Mar Vanrell for their comments and help in scripts and statistics.

Research funded by projects FFI2009-07648/FILO and CONSOLIDER-INGENIO 2010 Programme CSD2007-00012 (both awarded by the Ministerio de Ciencia e Innovación) and by project 2009 SGR 701 (awarded by the Generalitat de Catalunya).

Thank you!