Auditory Classification Images:

How noise can reveal the acoustic cues used in phoneme categorization.

Léo Varnet, *Gwendoline Trollé, Willy Serniclaes, Kenneth Knoblauch, Fanny Meunier, Michel Hoen*

Lyon Neuroscience Research Center, CNRS UMR 5292, Auditory Language Processing (ALP) research group, Lyon, France.

6th Workshop on Speech in Noise, Marseille, 09/01/2014

Decoding speech

Speech is a **complex code** (elements of the physical input → phonemes).

A major challenge for all psychologists in this field is to **crack the speech code** by finding which **acoustic cues** allow the listener to differentiate one phoneme from another.

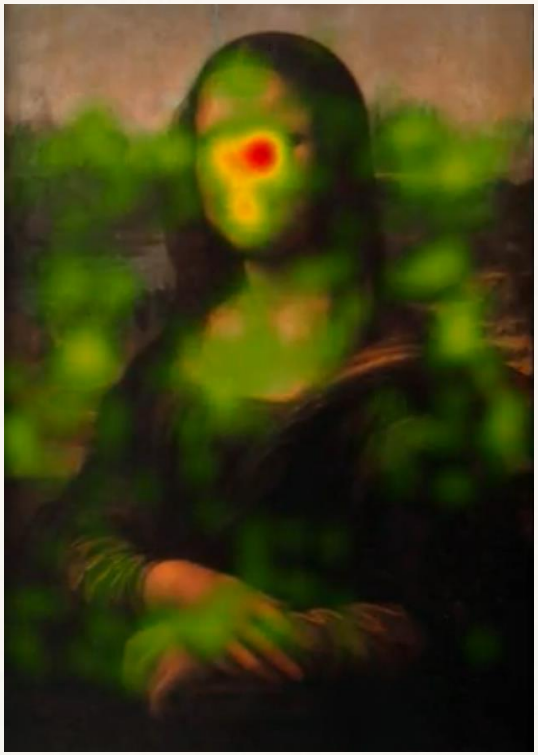
Why is this problem theoretically hard to solve?

- **Spectro-temporal complexity** of natural speech.
- The speech comprehension system shows very efficient **plasticity**.

The need for an "ear-tracker"

Our aim is to develop a new method to directly see where humans listen inside natural speech signals…

…like an **eye-tracker**:



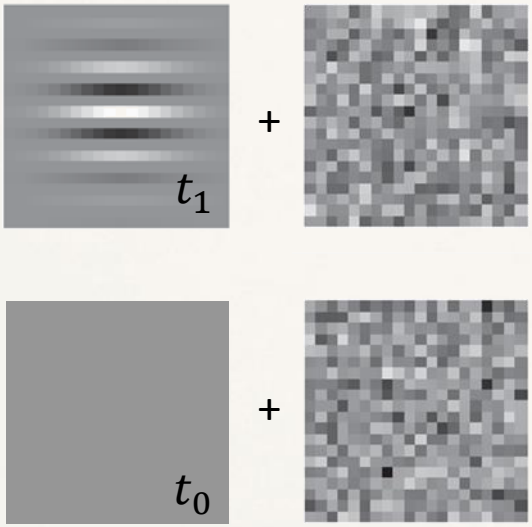
A solution could be provided by the technique of **Classification Image** (CIm).

Classification Image: brief description

Correlational technique *(Ahumada, 1971)* primarily used for applications in visual psychophysics.

Example : visual detection of a Gabor target in noise.

$s_i =$ t_0 or t_1 + noise N_i



Participant's categorization system (the "black box")

Response r_i

$r_i = 1\ (t_1)$

or

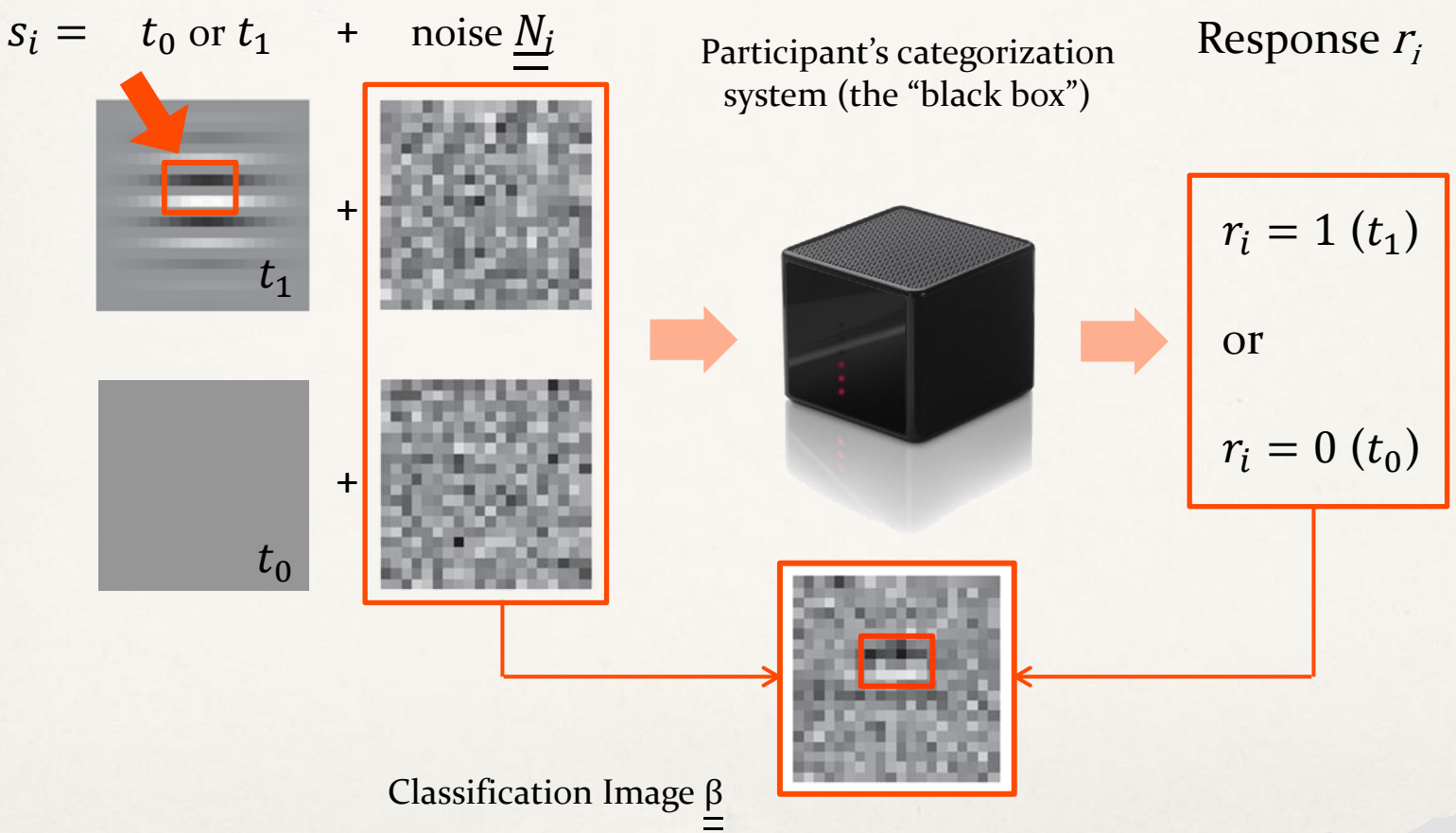
$r_i = 0\ (t_0)$

(Solomon, 2002)

Which information is used to detect whether the target was present or not ?

Classification Image: brief description

Correlation between the specific **noise field** in each trial and the **response** of the observer. The resulting correlation matrix shows how the presence of noise at each point interferes with the decision.



Aba/Ada experiment

Applying CIm to **the auditory modality** *(Varnet et al., 2013)*.

frontiers in
HUMAN NEUROSCIENCE

METHODS ARTICLE
published: xx December 2013
doi: 10.3389/fnhum.2013.00865

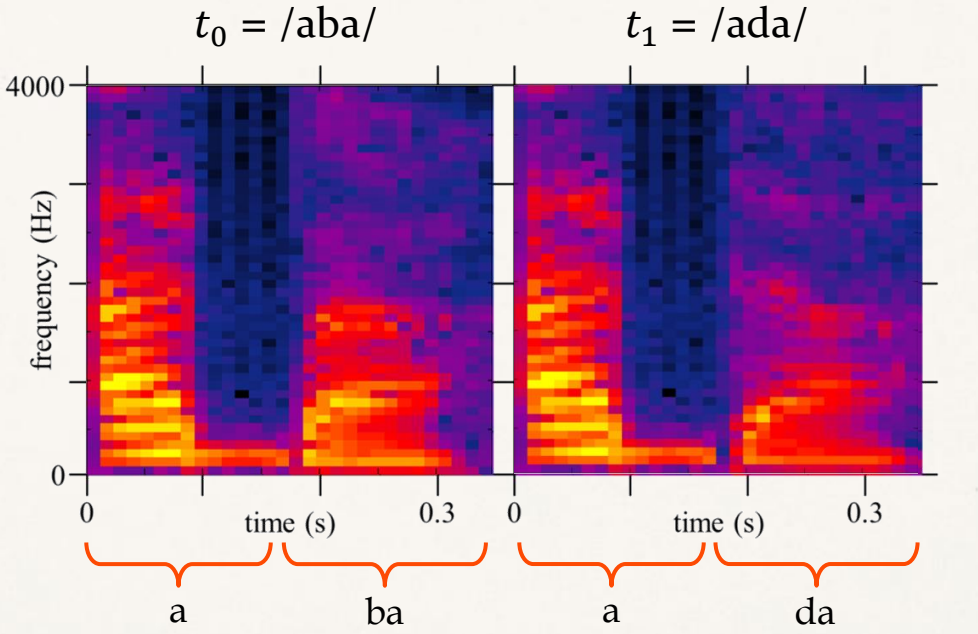
Using auditory classification images for the identification of fine acoustic cues used in speech perception

***Léo Varnet*[1,2]*, *Kenneth Knoblauch*[2,3], *Fanny Meunier*[1,2,4] *and Michel Hoen*[1,2]**

[1] *Neuroscience Research Centre, Brain Dynamics and Cognition Team, INSERM U1028, CNRS UMR5292, Lyon, France*
[2] *Université de Lyon, Université Lyon 1, Lyon, France*
[3] *Integrative Neuroscience Department, Stem Cell and Brain Research Institute, INSERM U846, Bron, France*
[4] *Laboratoire sur le langage le cerveau et la cognition, CNRS UMR5304, Lyon, France*

Materials…

Target : 2 speech sounds (t_0 = /aba/ and t_1= /ada/) obtained by concatenating the same utterance of /a/ with two single utterances of /ba/ and /da/ (power-normalized).



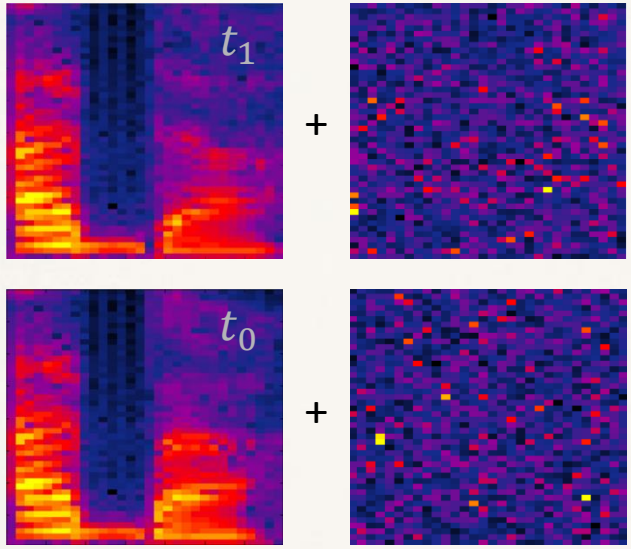
Stimuli: Target sounds in an additive Gaussian noise.

Task: Indicate whether the target was /aba/ or /ada/.

The SNR was adapted continuously to ensure a correct response rate of 75%.

…& Methods

$s_i =$ t_0 or t_1 + noise $\underline{N_i}$



Participant's categorization system (the "black box")

Response r_i

$r_i = 1\ (t_1)$

or

$r_i = 0\ (t_0)$

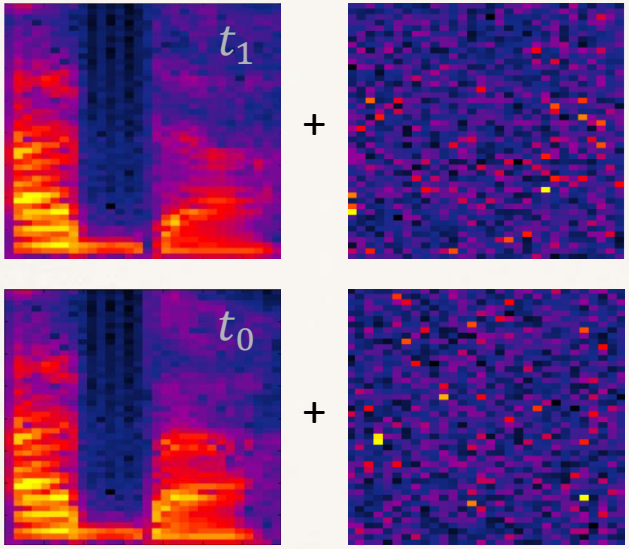
Two major differences:

1) The analysis is based not on the Gaussian noise but on the **time-frequency representation of the noise.**

2) The Auditory CIm will require more trials than the Visual CIm to be computed accurately.

…& Methods

$s_i =$ t_0 or t_1 + noise $\underline{N_i}$



Participant's categorization system (the "black box")

Response r_i

$r_i = 1\ (t_1)$

or

$r_i = 0\ (t_0)$

Generalized Linear Model (GLM) with **smoothness priors** framework**:**

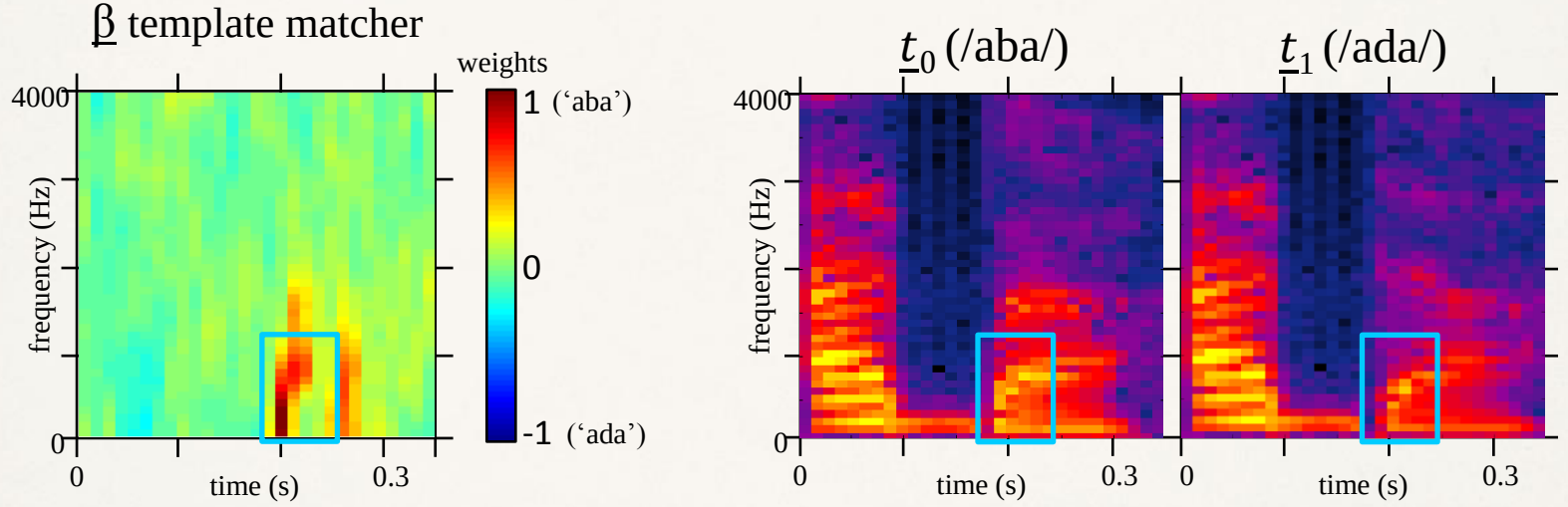
$$P(r_i = 1) = \Phi(\underline{N_i}^T \cdot \underline{\beta} + \beta_{t_i})$$

The parameters (the β's) are fitted by a **maximum a posteriori estimation** algorithm.

Modeling an ideal template-matcher

What would be the results if the participants performed the task linearly by comparing the input stimulus with templates of the two targets and choosing the most similar?

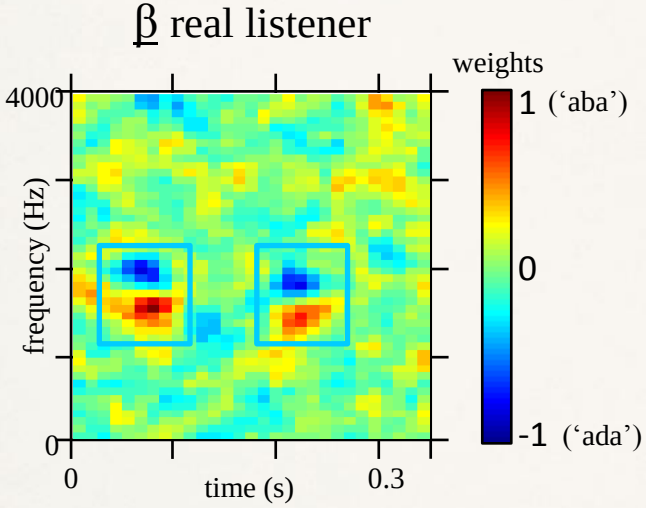
We simulated an experiment performed by an **ideal template-matcher**.

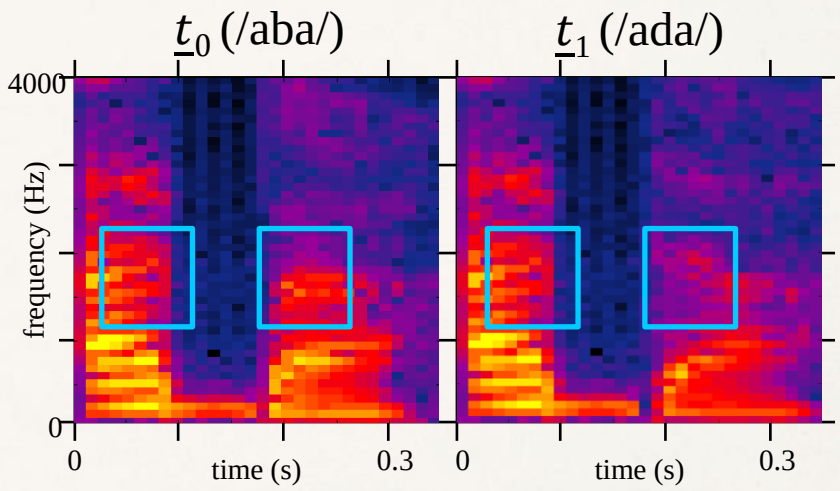


The ideal template-matcher follows the **optimal strategy**: taking into account only **the region where the targets differ most** in terms of energy.

Classification Images: real listener

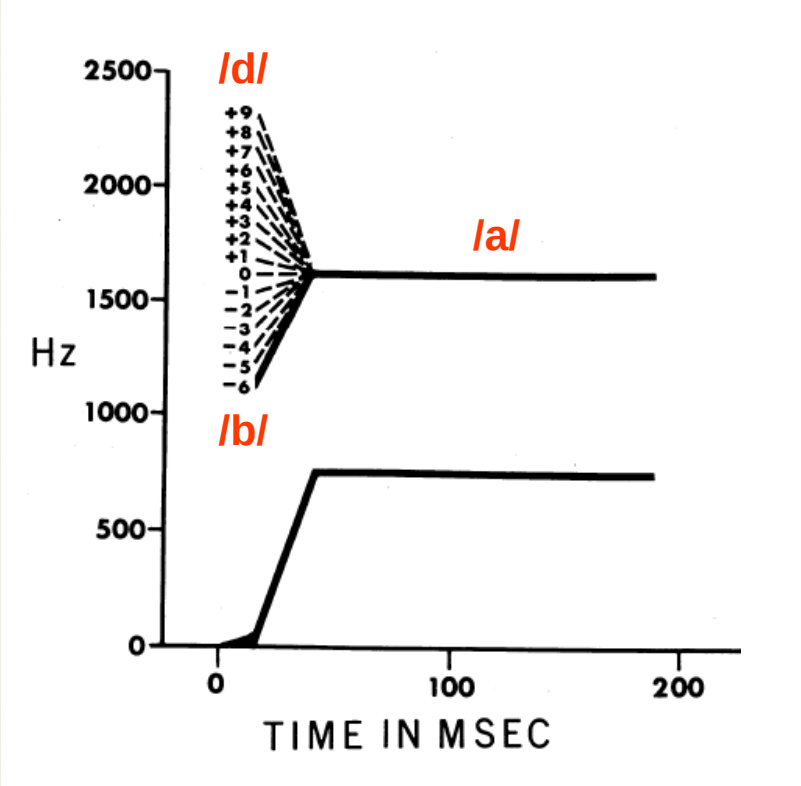
- It does not look like the CIm of the ideal template-matcher…
- **Two similar patterns** (a cluster of positive weights below a cluster of negative weights) at two precise time-frequency locations.

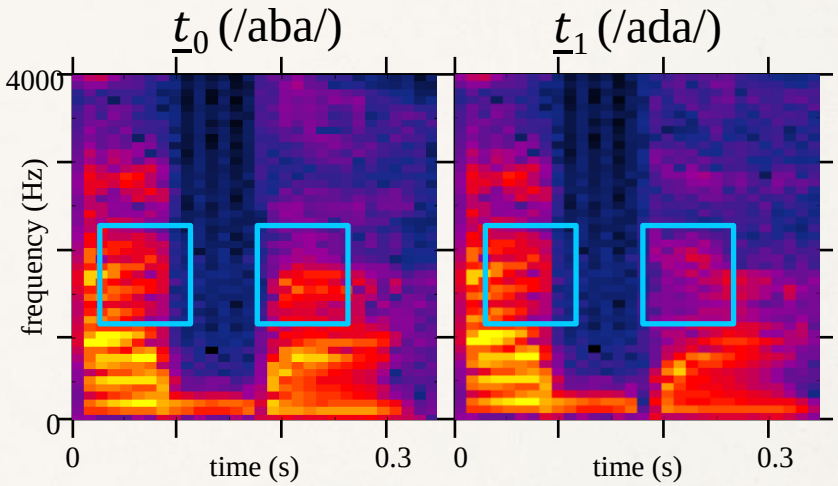




The critical time-frequency locations exactly match the **F2 transitions** !

Classification Images: real listener

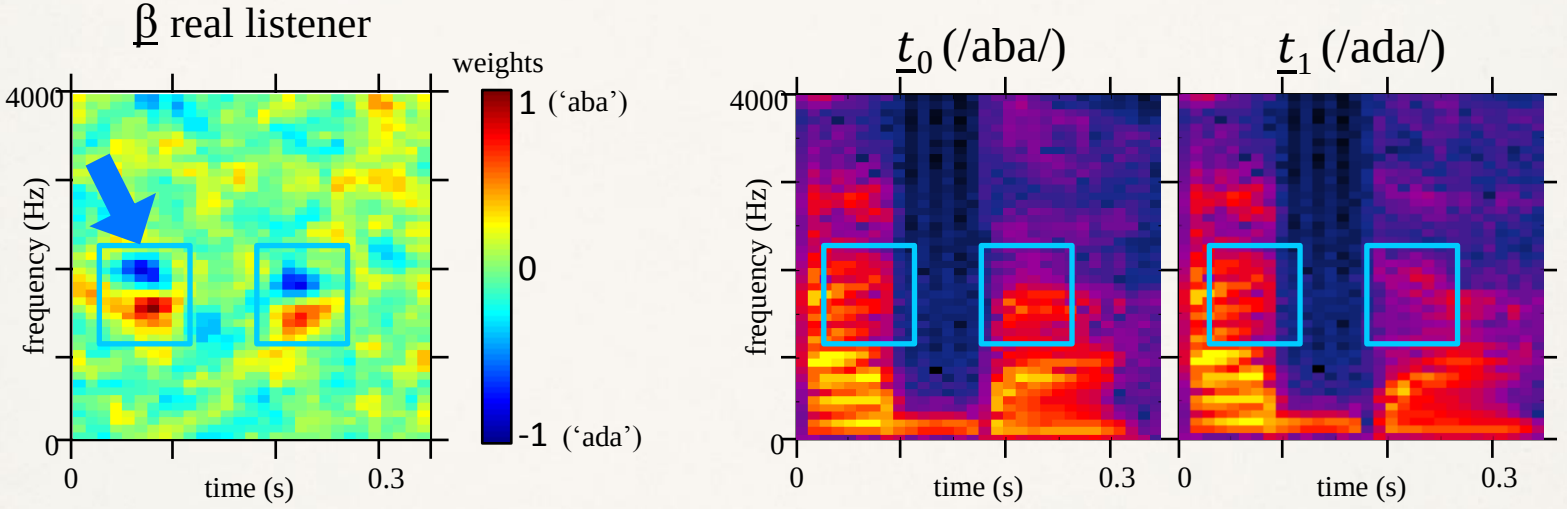




Liberman (1954): The second formantic transition is a key for classifying phonemes into /b/ or /d/.

Classification Images: real listener

- This measurement of the F2 onset frequency by the auditory system is a **relative estimation**.
- The real listener also used information from the first syllable.

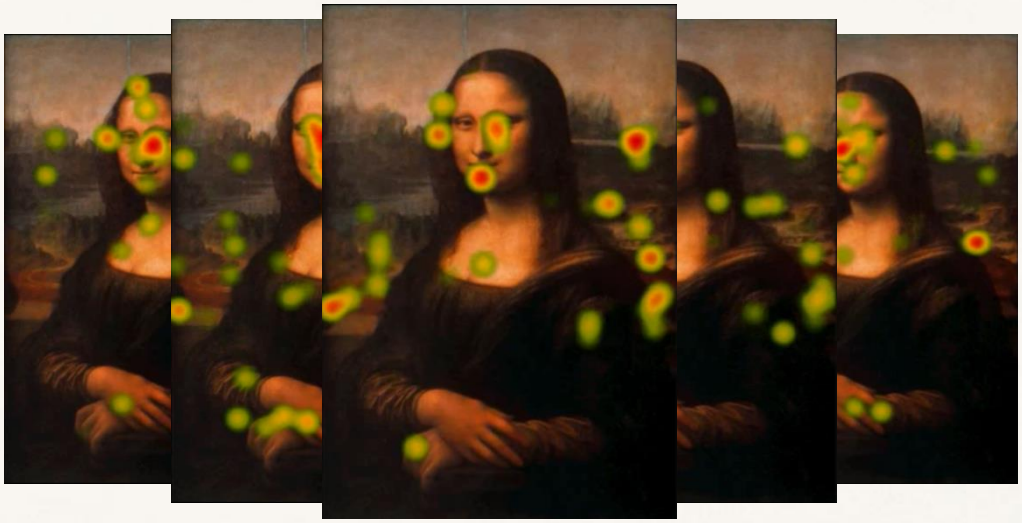


- This could allow us to **compensate for variability** in speech production…

The participant extracts information from the first syllable, even though this region actually contains **no useful acoustic cue** for performing the task.

What remains to be done ?

- The ear-tracker works fine: **We can see where people listen to**.



- For future applications, we will need **more participants** and a **statistical assessment at the group level**.
- Gathering images from several participants will **reduce the number of trials needed** from each participant.

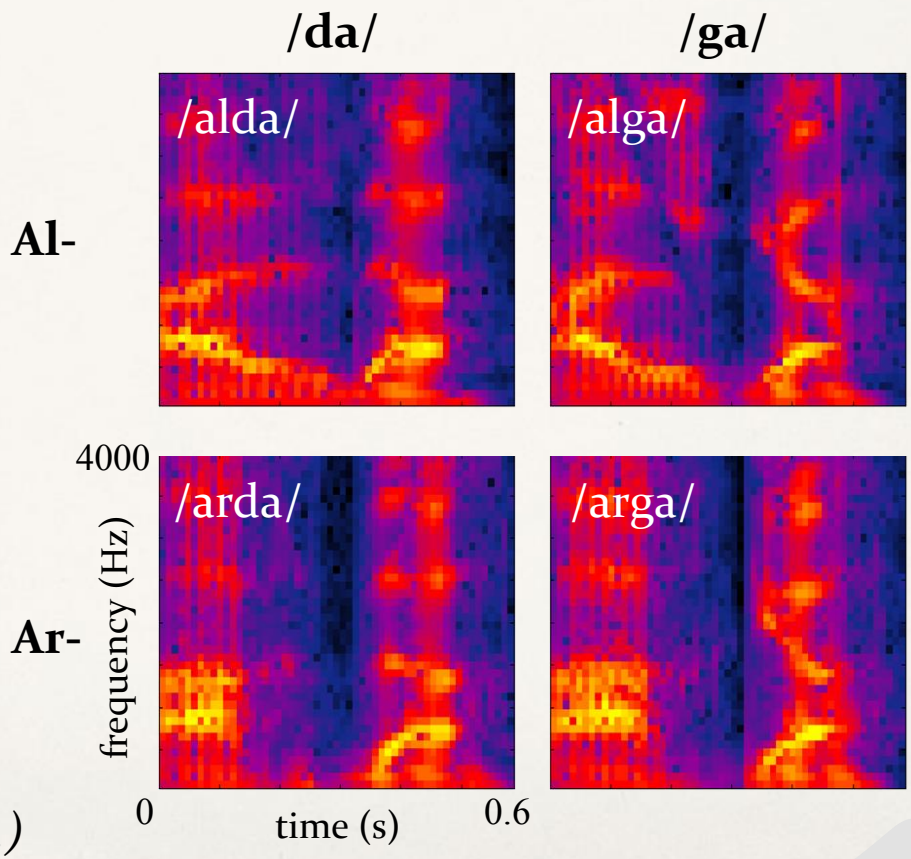
2nd study : Mann's experiment

Signals: **4 VCCV sequences**: /alda/, /alga/, /arda/, /arga/ (natural speech productions, equated in duration and rms)

Task: Indicate whether the last syllable was **/da/ or /ga/**.

This particular situation is a classic example of **how phonetic context influences phonemic categorization**. (*Mann, 1980; Holt, 2006; Fowler, 2006…)*

Could Auditory CIs give us some insights on how listeners deal with phonetic context ?



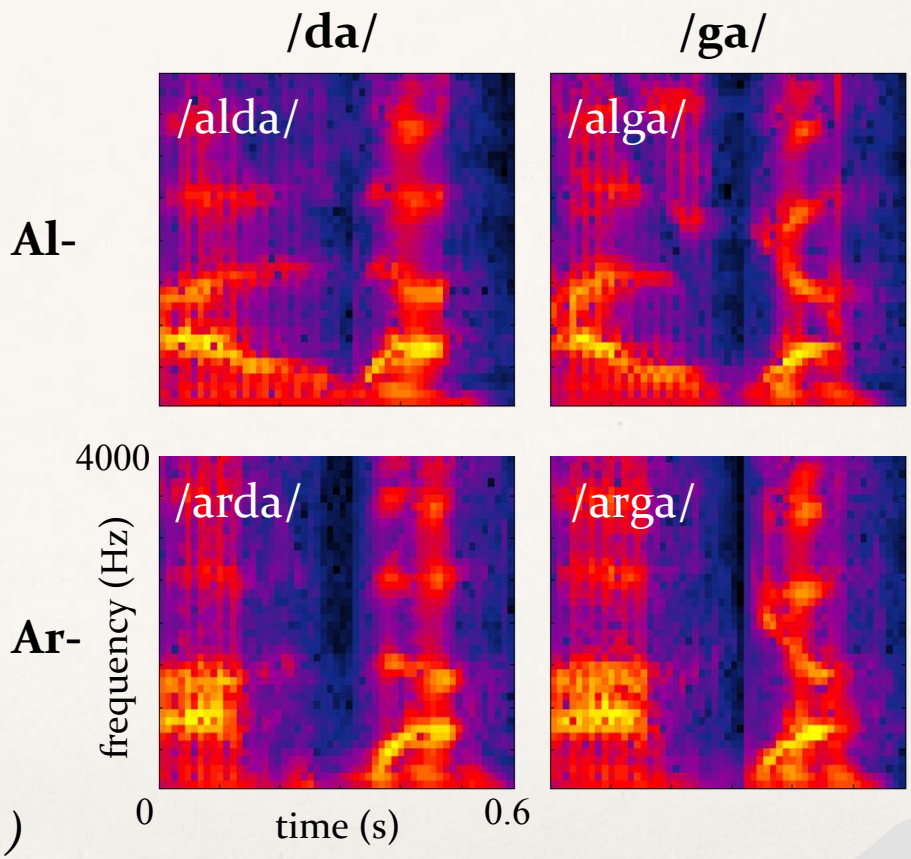
(Varnet et al., in prep.)

2nd study : Mann's experiment

Stimuli: Targets in gaussian noise. SNR was adapted continuously to ensure a correct response rate of 79% ("3-down 1-up" staircase algorithm).

Participants: 16 native French speakers, aged 19-35 (12 women + 4 men, no hearing or language disorders, non-musicians).

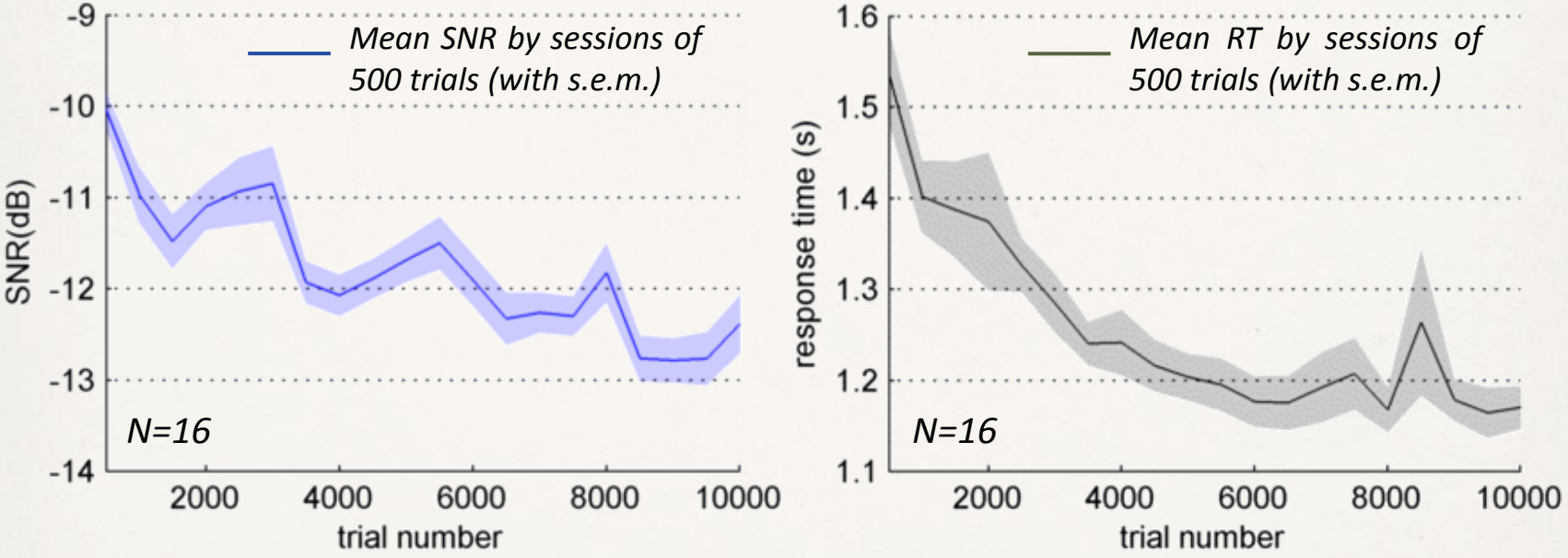
Each participant completed a set of **10.000 trials** in random order (20 sessions of 500 trials over 4 days).



(Varnet et al., in prep.)

General performances

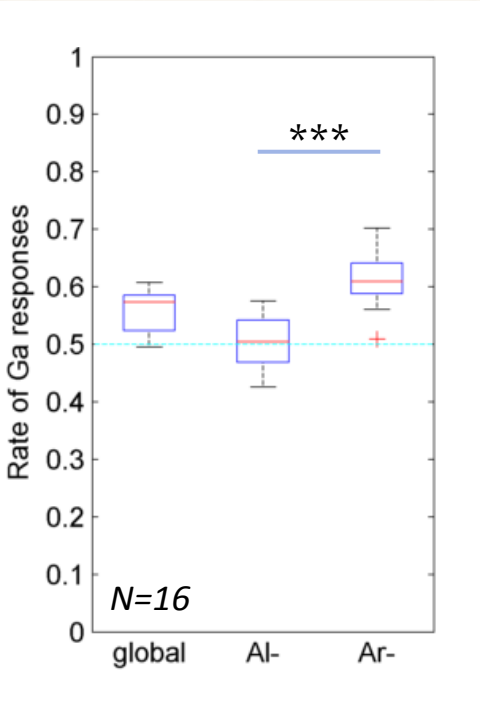
- Mean correct response rate of ~80% for each participant.
- SNR decreases over the course of the experiment, as well as response time.



Mean evolution of SNR and Response Time over the course of the experiment.

General performances

- Mean correct response rate of ~80% for each participant.
- SNR decreases over the course of the experiment, as well as response time.
- Slight but significant **bias** of all participants **towards response "ga".**
- Participants are biased **only in context Ar-**.

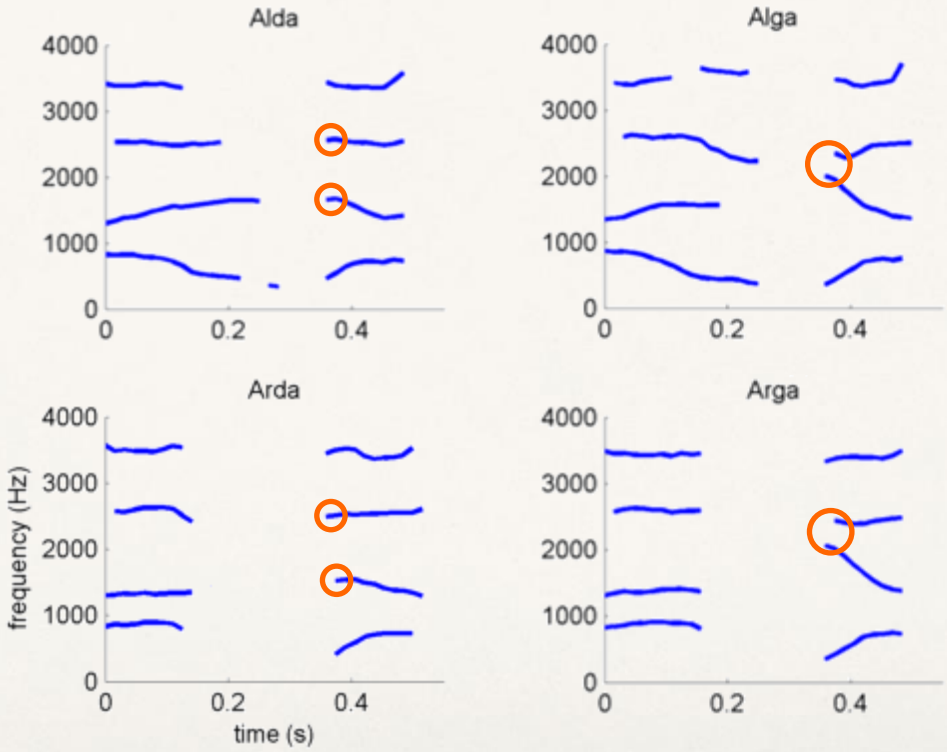


Mean bias in each condition.

- **Phonetic context** may influence the categorization of the target (phonotactic effect, compensation for coarticulation…).
- **Our particular utterance** of "Arga" may simply be more distinctly produced than "Arda".

Mann's experiment

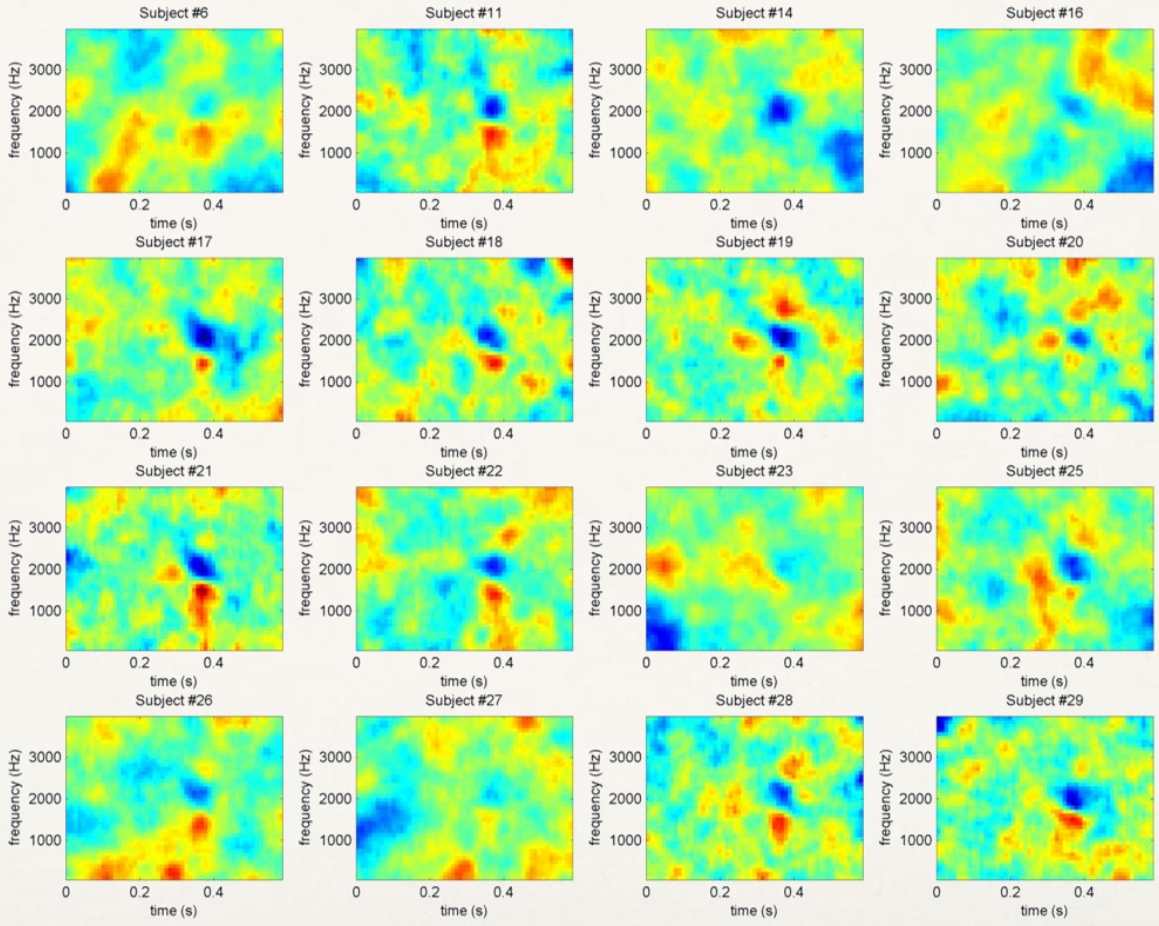
- Natural speech utterances: we didn't control for formant positions between our 4 stimuli.
- The **onsets of the 2nd and 3rd formants** would be critical for correct categorization of "da" and "ga" *(Stephens & Holt, 2003 ; Viswanathan et al., 2010…)*.



Formant trajectories for our 4 stimuli with possible acoustic cues for the categorization.

Results: acoustic cues

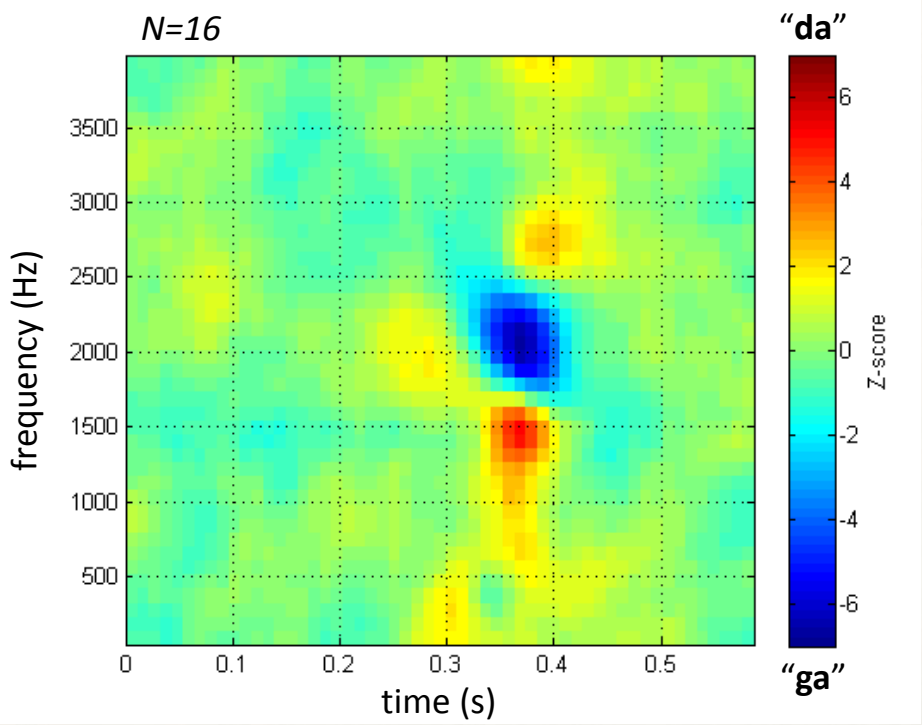
- Similar pattern of weights for all participants, despite some variability.



Individual Classification Images for the 16 participants.

Results: acoustic cues

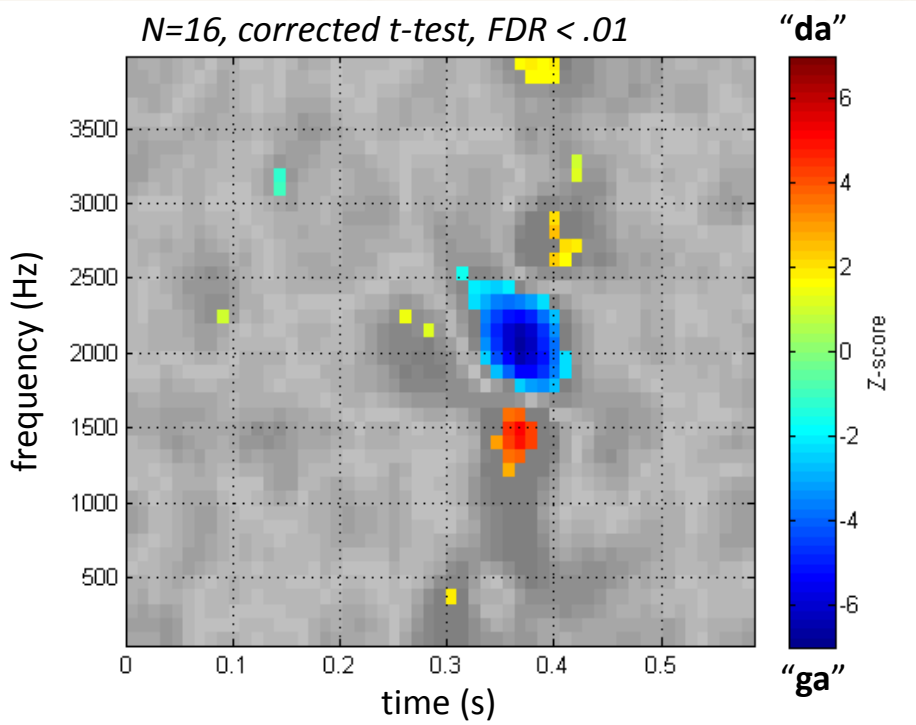
- Similar pattern of weights for all participants, despite some variability.
- One negative acoustic cue surrounded by two positive acoustic cues.



Mean Classification Image over all participants

Results: acoustic cues

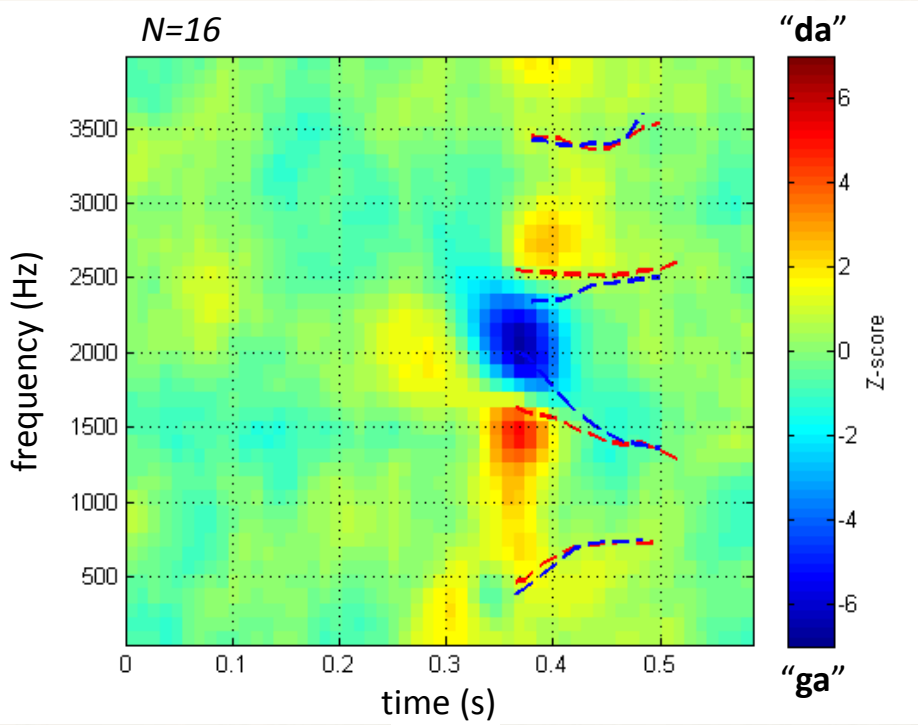
- Similar pattern of weights for all participants, despite some variability.
- One negative acoustic cue surrounded by two positive acoustic cues.



Mean CIm over all participants, t-test against 0 with FDR correction

Results: acoustic cues

- Similar pattern of weights for all participants, despite some variability.
- One negative acoustic cue surrounded by two positive acoustic cues.
- The **2nd and 3rd formantic transitions** are critical cues for this task.
- **No anticipatory cue** in the 1st syllable (contrary to the Aba/Ada experiment): the categorization already relies on the relative positions of 2 formants.



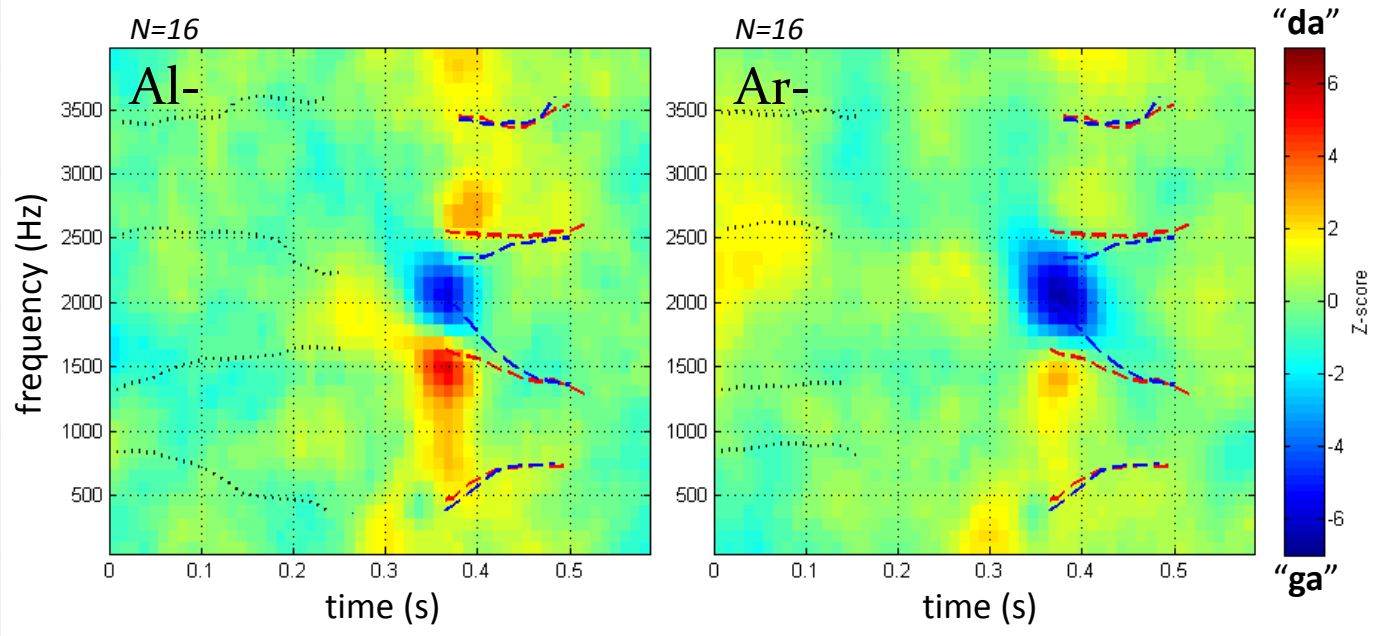
Mean Classification Image over all participants

Formant trajectories:

- - - - - *"da"*
- - - - - *"ga"*

Results: context effects

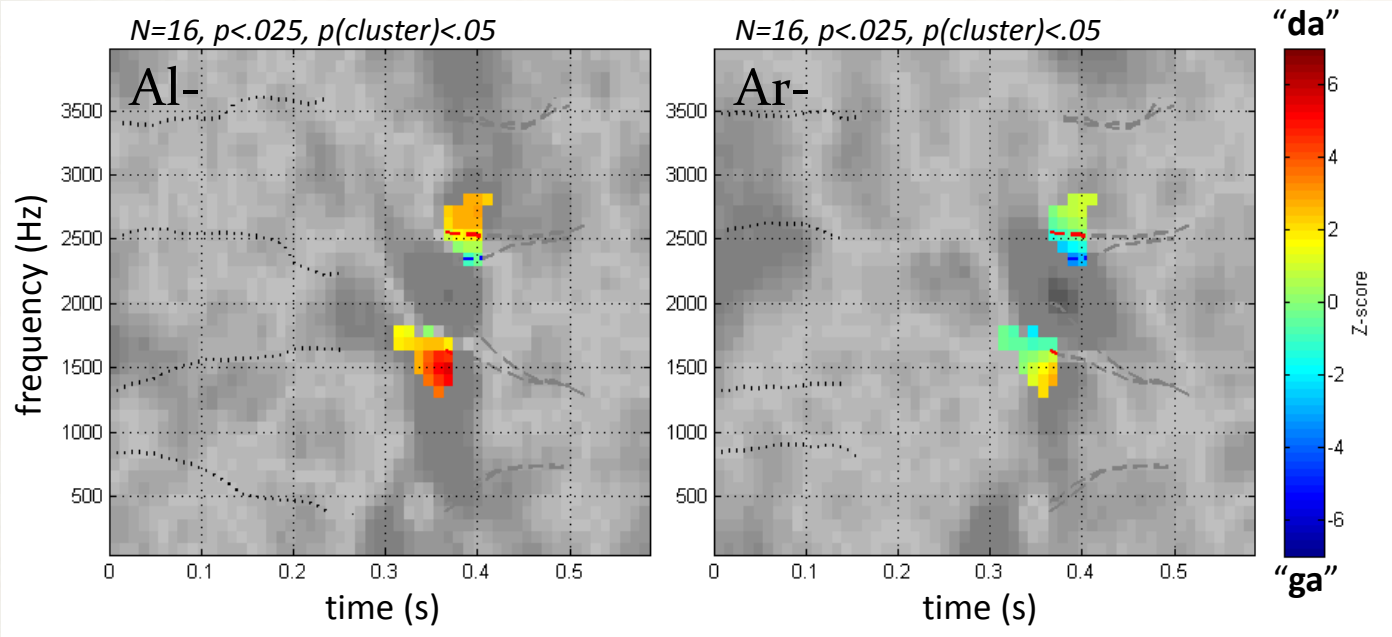
- Comparison context **Al-** and **Ar-**: positive cues are significantly decreased.



Mean Classification Image over all participants, calculated for conditions Al- and Ar-

Results: context effects

- Comparison context **Al-** and **Ar-**: positive cues are significantly decreased.

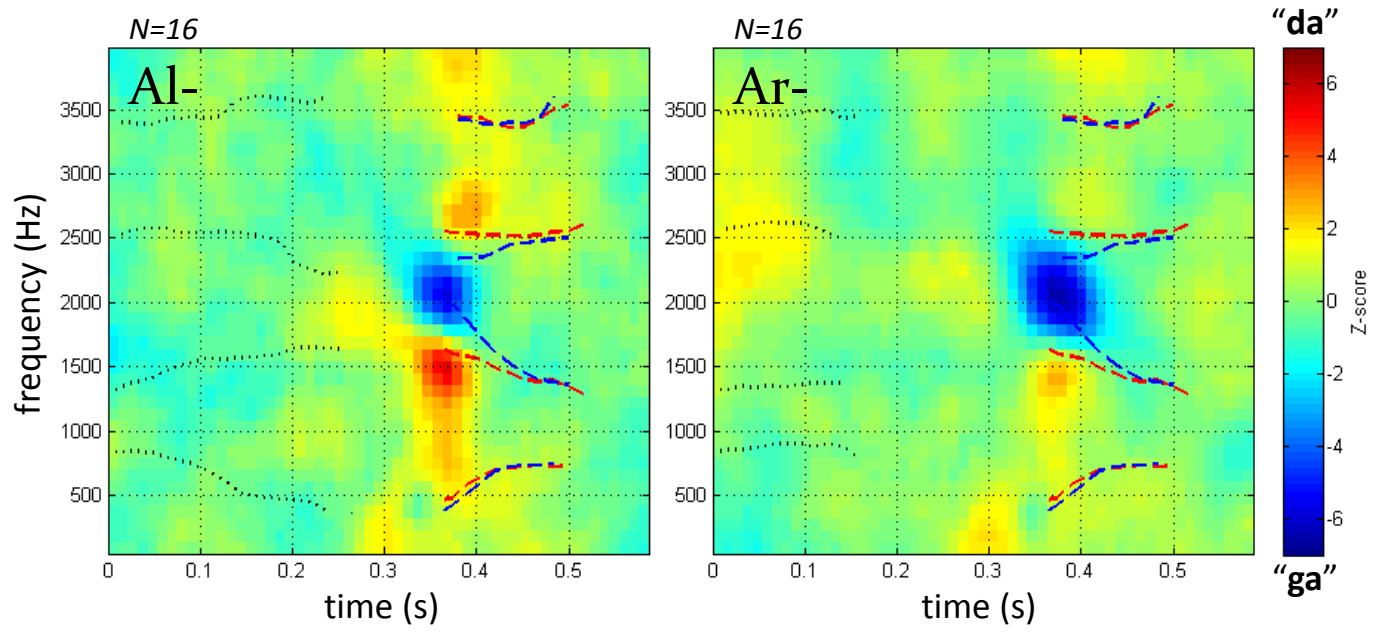


Mean Classification Image over all participants, calculated for conditions Al- and Ar-
Significant difference between conditions (cluster-based nonparametric test)

Results: context effects

- Comparison context **Al-** and **Ar-**: positive cues are significantly decreased.
- In **context Ar-**, participants are **more sensitive to the "ga" cue** than to the "da" cues ⇒ consistent with the observed bias towards "ga" in context Ar-.
- True context effect ? Or the central cue is simply more salient in "Arga" ?

CIm are precise enough to **track fine modifications** between listening conditions.

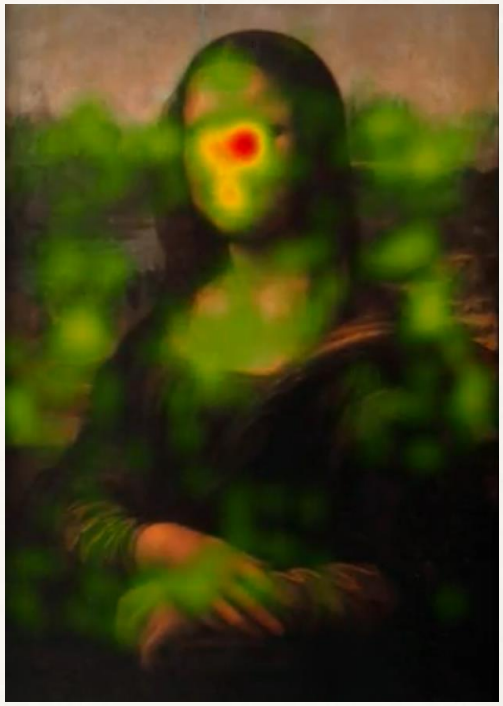


Mean Classification Image over all participants, calculated for conditions Al- and Ar-

Where do we stand ?

?

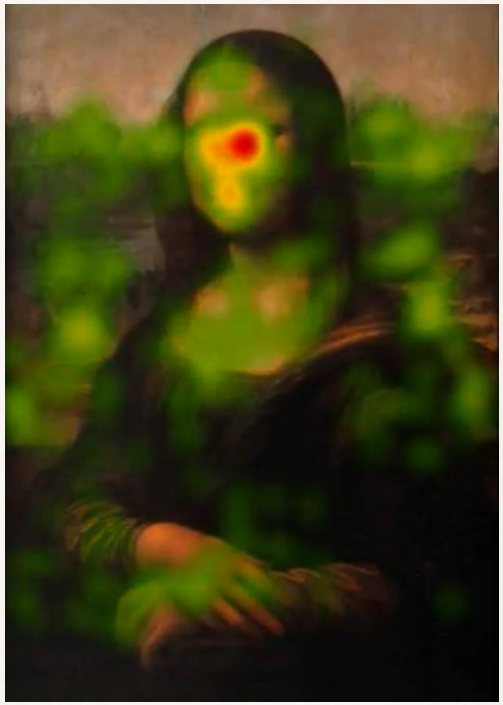
- We can see where people listen to.
- Identify **functional cues**.
- CIs both at the **individual** and **group** levels.
- **Natural speech** stimuli, instead of synthetic speech.



Where do we stand ?

?

- We are **far from single-trial resolution**.
- The Classification Image method requires the stimuli to be presented with **low SNR**.
- We are forced to use a **limited number of stimuli**, and participants are restricted to a small number of possible answers.



The Auditory Classification Image technique is a promising psychoacoustic method giving us an **insight into the mechanisms of speech perception**.

Thank you for your attention !