Assessment of individual listening strategies in amplitude-modulation detection and phoneme categorisation tasks

Alejandro Osses, Christian Lorenzi, Léo Varnet

ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

Assessment of individual listening strategies in amplitude-modulation detection and phoneme categorisation tasks

Alejandro OSSES VECCHI, Christian LORENZI, Léo VARNET

Laboratoire des systèmes perceptifs, Departement d'Études Cognitives, École normale supérieure, Université Paris Sciences & Lettres, Centre National de la Recherche Scientifique, 75005 Paris, France

alejandro.osses@ens.psl.eu, christian.lorenzi@ens.psl.eu, leo.varnet@ens.psl.eu

ABSTRACT

Auditory reverse correlation (revcorr) is an experimental paradigm that allows researchers to reveal the acoustic features that are used as cues by listeners in an auditory task. The paradigm relies on a stimulus-response model, fitted using a penalised logistic regression, to produce a time-frequency matrix of decision weights called auditory classification image (ACI). An ACI provides thus a map of the participant's listening strategy in a given task. In this study, we will present results obtained by two participants in a series of auditory revcorr experiments. In all experiments participants had to indicate which of two possible target sounds were presented. The two target sounds were: (1) modulated and non-modulated tones, or (2) /aba/ and /ada/ speech samples uttered by different speakers. A key ingredient in the revcorr method is the presentation of target sounds embedded in an additive background noise at a signal-to-noise ratio such that the additive noise can affect the participants' performance. All experiments are implemented in our in-house fastACI toolbox, which offers a ready-to-use solution for setting up, running, and analysing an auditory revcorr experiment.

Keywords: Reverse correlation, auditory classification images, amplitude modulation, speech perception, acoustic cues

1 INTRODUCTION

Auditory reverse correlation (revcorr) is an experimental paradigm that allows researchers to reveal the acoustic features used as cues by listeners in an auditory task. Originally proposed by Ahumada and Lovell [1], the method has become very popular in the psychoacoustic community during the last decade, with applications to loudness perception, tone-in-noise detection, modulation perception, timbre judgement, phoneme-in-noise perception, sentence recognition, and prosody perception (see http://dbao.leo-varnet.fr/2020/12/03/a-visual-compendium-of-auditory-revcorr-studies/ for a non-exhaustive review). In the present study, we will focus on a specific version of the auditory revcorr paradigm particularly suited for exploring simple auditory categorisation tasks and called auditory classification image (ACI).

In a typical ACI experiment, participants are asked to discriminate two sounds that are repetitively presented in a background noise. This paradigm, which corresponds to a very simple form of sound categorisation, is illustrated in Figure 1 for a phoneme-in-noise task with two words of the structure vowel-consonant-vowel, or logatomes, that differ in only one single phonetic feature [2]. In some trials, the random distribution of noise will mask crucial acoustic characteristics of the target, resulting in an incorrect response of the participant. A statistical estimation of the stimulus-percept relationship allows to reveal the time-frequency (T-F) regions where the presence of noise affects the participant's decision in a systematic way, i.e., the regions corresponding to the acoustic cues on which they relied to successfully discriminate the stimuli. The ACI method can therefore be described as a perceptual imaging technique as it provides a direct visualisation of the listening strategy used by a participant during the task. This novel methodology is based on purely behavioural data (no neuroimaging data) and offers an unprecedented insight into the mechanisms at play at the acoustic-phonetic interface.

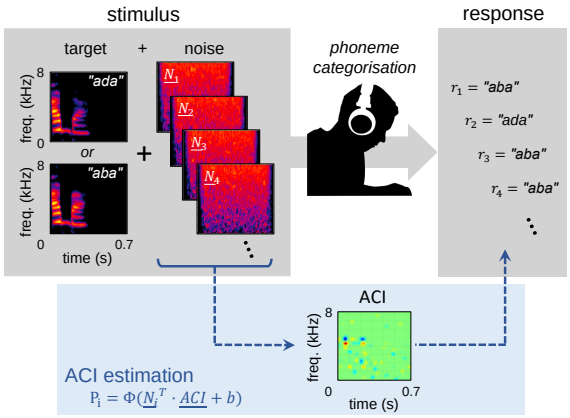

Figure 1. Schematic diagram of a typical ACI experiment (grey) and ACI estimation (blue), as evaluated by Varnet *et al.* [2]. In each trial, one speech target (here, /aba/ or /ada/) is chosen at random and embedded in an additive white noise. After each stimulus presentation, the participants are asked to indicate which target they heard. The rationale behind an ACI experiment is to find the systematic relationship between the T-F noise representation and the corresponding response of the listener, on a trial-by-trial basis. To do so, a prediction of the binary decision (the choice of /aba/ or /ada/) is obtained from a statistical (GLM) estimation fitted using a penalised logistic regression (see Section 3.4.2 for details). The resulting matrix of weights, the ACI, reveals the T-F regions where the presence of noise misled the participant in a systematic way, indicating the location of the acoustic cues on which the participant relied during the experiment.

The fastACI toolbox developed in our group and freely available at https://github.com/aosses-tue/fastACI, offers a ready-to-use solution for setting up, running, and analysing an ACI experiment [3]. In particular, it allows to reproduce all our previous auditory revcorr studies on modulation perception [4] and phoneme perception [2], [5], [6]. In the present study, we re-analyse the data for a selection of previous experiments using a single analysis pipeline to obtain a longitudinal view of the listening strategy of two participants in a series of four auditory experiments. Our goal is not so much to replicate or extend previous revcorr studies but to rather explain and illustrate the revcorr paradigm on auditory and phonetic tasks. Therefore, our study is primarily didactic in nature. As part of our dissemination strategy, we recently presented portions of this study in oral form at a local symposium [7].

2 METHODS

This study aggregates data from several experiments performed by the same two participants, labelled as SA and SB. All experiments were based on single-interval trials containing one of the two target sounds presented in noise using either a white-noise or a speech-shaped-noise (SSN) masker: (1) **ABDA13** replication of the /aba/-/ada/ discrimination in white noise from Varnet *et al.* [2] for which we adopted their same female speech productions and noise waveforms but employed an updated experimental protocol, (2) **ABDA21** /aba/-/ada/ discrimination in SSN [5] using the female speech productions from [2], (3) **ABDA22** /aba/-/ada/ discrimination in white noise, similar to ABDA13 but using two male speech productions [8], [9], and (4) **MOD22** amplitude-modulation (AM) detection in white noise [4]. The specific parameters used in each experiment are listed in Table 1 and the auditory targets are shown in Figure 2. All experiments can be reproduced within the fastACI toolbox (see Appendix A.1 for further details).

2.1 Experimental protocol

The trial structure of a typical ACI (revcorr) experiment, exemplified for ABDA13, is shown in Figure 1. The same protocol was followed in ABDA21, ABDA22, and MOD22, but using different targets and maskers, as indicated in Table 1. All four experiments were based on a discrimination task in noise with two equally-likely targets presented through headphones. These simple "categorisations" consisted of N trials, with each trial having one target-in-noise interval to which the participant had to indicate one of the two possible answers.

Table 1. List of parameters used in each experiment. The adaptive procedures used in different experiments targeted the same overall performance score of 70.7%. Further details are given in the body text. Abbreviations: WN = white noise; SSN = speech-shaped noise; N = total number of collected trials.

					Step size (dB)			
Exp. name	Masker	Targets (gender speaker)	N	Staircase rule	up	down	Roving	Ref.
ABDA13	WN	/aba/ or /ada/ (female)	5000	transformed 1-up 2-down	1	1	no	[2]
ABDA21	SSN	/aba/ or /ada/ (female)	5000	weighted 1-up 1-down	1	0.41	no	[2], [5]
ABDA22	WN	/aba/ or /ada/ (male)	4000	weighted 1-up 1-down	1	0.41	$\pm$ 2.5 dB	[8]
MOD22	WN	modulated or unmod. tone	3000	transformed 1-up 2-down	1	1	no	[4]

Participants were not allowed to repeat trials, but we provided feedback about the correctness of their responses. The level of the noise was fixed to 65 dB SPL (64 dB SPL in MOD22) while the level of the target (or the modulation depth in MOD22) started at $\mathrm{SNR}_{\mathrm{init}}$=0 dB (or $\mathrm{m}_{\mathrm{depth,init}} = -1$ dB) and was adapted during the entire experiment using a staircase procedure to yield a 70.7% correct performance threshold with either a transformed [10] or a weighted up-down rule [11]. In ABDA22, a small roving was applied to the presentation level of the trial to discourage the use of loudness cues (see Table 1). Due to the large number of required trials, the experiments were organised in short blocks of less than 20 minutes (containing 400 or 500 trials) across several days. Each experimental session lasted typically two hours and contained between 4 and 6 blocks. All the noises were stored in the test computer together with the corresponding participants' responses.

3 Target stimuli

In this section we present a description of the target sounds used in the experiments. These targets were always presented together with a background noise, or masker, that was expected to detriment the properties of the target stimuli (summarised in Table 2, see below). Our description does not focus on explaining how the acoustic properties of the noises interact with those of the targets—the interested reader is referred to [4], [8], [9], although the perceptual effects of that interaction is what is estimated with the revcorr method.

3.1 ABDA13 and ABDA21

The /aba/ and /ada/ sounds were productions from a female speaker, recorded for the original study by Varnet *et al.* [2]. The /aba/ and /ada/ recordings had an average fundamental frequency $f_0 = 222.2$ Hz (std= 32 Hz) and $f_0 = 197.6$ Hz (std= 28.6 Hz), respectively. For the replication of ABDA13 and for ABDA21, the speech sounds had a duration of 0.84 s, that is, 0.69 s of the original recordings, zero padded at the beginning and at the end by 0.075 s to match the total duration of the noises (that had 0.075-s up/down ramps). The level of the zero-padded sounds was 65 dB SPL. See Table 2 for further details and Figure 2 for a graphical representation of the two speech samples.

3.2 ABDA22

The /aba/ and /ada/ sounds were natural male productions (native French speaker S43M) taken from the OLLO database [12]. The sounds were pre-processed to have equal duration and similar acoustic energy in the first and second syllables [8]. The /aba/ and /ada/ recordings had an average fundamental frequency $f_0 = 110.5$ Hz (std= 4.6 Hz) and $f_0 = 110$ Hz (std= 4.7 Hz), respectively. The resulting sounds had a duration of 0.85 s, after zero padding for the same purposes as indicated for ABDA13. The total level of the sounds was 65 dB SPL. See Table 2 for further details and Figure 2 for a graphical representation of the two speech samples.

3.3 MOD22

The reference sound was a pure tone of frequency $f = 1000$ Hz with a duration of 0.75 s, including cosine ramps of 0.075 s at the beginning and end of the stimulus. The target sound was a sinusoidally amplitude-modulated version of the reference with a rate $f_{\mathrm{mod}} = 4$ Hz and initial phase of $3\pi/2$, i.e., starting at a modu-

Table 2. Characterisation of the target stimuli, the vowel-consonant-vowel (VCV) words used in the ABDA experiments. The stimuli are available on the fastACI toolbox repository [3], under the folder `Stimuli`. The fundamental frequency (f_0) and formants (F_1 to F_4) were extracted using the Praat software. The signal levels were obtained as root-mean-square values over segments excluding silent fragments of the waveforms. The level of the waveforms (with no silence exclusion) over the total length of the waveforms equals 65 dB SPL. The values in parentheses represent the standard deviation of the corresponding frequency (f_0 or F_1–F_4) in Hz.

	Speech sound /aba/			Speech sound /ada/		
	First syllable	Second syllable	VCV word	First syllable	Second syllable	VCV word
ABDA13 & ABDA21	`../Stimuli/varnet2013/Aba.wav`			`../Stimuli/varnet2013/Ada.wav`		
f_0 (Hz)	252.5 (7.9)	191.9 (4.3)	222.2 (32.0)	252.5 (7.9)	197.6 (5.7)	232.6 (28.6)
F_1 (Hz)	897.4 (84.6)	796.2 (64.9)	846.8 (89.4)	897.4 (84.6)	682.8 (82.1)	819.3 (134.3)
F_2 (Hz)	1784.8 (35.5)	1605.4 (29.5)	1695.1 (98.2)	1784.9 (35.5)	1849.1 (33.8)	1808.2 (46.4)
F_3 (Hz)	2775.6 (33.9)	3035.7 (144.2)	2905.7 (168.3)	2775.6 (33.9)	2900.0 (32.9)	2820.9 (70.4)
F_4 (Hz)	4311.4 (141.5)	3674.1 (103.9)	3992.8 (351.5)	4311.4 (141.4)	4074.0 (50.7)	4225.1 (164.7)
Level (dB SPL)	70.9	65.4	67.7	71.7	64.0	67.7
Start-end time (ms)	50–150	280–420	50–420	50–150	280–420	50–420
ABDA22	`../Stimuli/Logatome/S43M_ab_ba.wav`			`../Stimuli/Logatome/S43M_ad_da.wav`		
f_0 (Hz)	112.8 (1.3)	102.6 (1.1)	110.5 (4.6)	112.1 (2.6)	102.8 (0.6)	110.0 (4.7)
F_1 (Hz)	732.9 (120.4)	751.5 (47.8)	744.0 (82.2)	679.3 (123.8)	715.0 (82.7)	702.0 (98.2)
F_2 (Hz)	1361.8 (54.7)	1314.4 (18.6)	1333.4 (43.2)	1461.9 (22.6)	1466.9 (149.5)	1465.1 (118.4)
F_3 (Hz)	2431.6 (36.9)	2507.4 (19.7)	2477.0 (46.7)	2389.0 (41.3)	2443.1 (56.6)	2423.4 (57.1)
F_4 (Hz)	3523.8 (109.8)	3687.4 (89.2)	3622.0 (125.7)	3597.8 (127.5)	3638.0 (43.8)	3623.4 (83.7)
Level (dB SPL)	68.7	65.6	66.5	68.5	65.8	66.5
Start-end time (ms)	50–150	280–550	50–550	50–150	280–550	50–550

lation dip. The modulated tone had an adaptive modulation depth m_{depth}. The reference and target sounds were adjusted to have a level of 54 dB SPL such that the noisy trials were at a fixed SNR of -10 dB (i.e., with a noise level of 64 dB SPL) resulting in a presentation level of 65 dB SPL.

3.4 Analysis

3.4.1 Time-frequency noise representations

ABDA experiments

Following the same rationale as in previous studies [5], each noise waveform was converted into a T-F representation. A Gammatone-based filter bank was used with 64 bands equally spaced in the ERB scale [13] between 40 and 8000 Hz (toolbox option `TF_type='gammatone'`, see Listing 2 in Appendix A.2). The 64 band-passed signal was then used as input to a simplified model of inner-hair-cell envelope processing [14, their Sec. II.3], [15, their Sec. 2.4]. The obtained T-F matrix was down-sampled to a temporal resolution of 0.01 s obtaining matrices $\underline{\underline{N_i}}$ (named `Data_matrix` within fastACI) that have 84-by-64 (ABDA13 and ABDA21) or 86-by-64 elements (ABDA22). $\underline{\underline{N_i}}$ was subsequently reshaped into a vector $\underline{N_i}$ with 5376-by-1 and 5504-by-1 elements, respectively.

MOD22 experiment

The same processing was applied to obtain T-F representations in MOD22 with the only exception that instead of using the noise waveforms alone, the non-modulated tone used in the experiment was added to each noise. Varnet and Lorenzi argued that this processing is needed to discard the task-relevant information in the stimuli [see 4, their Sec. II**D**]. We decided to keep the nomenclature $\underline{N_i}$ for the T-F "tone-plus-noise" vector of this experiment to be able to adopt the same matrix notation across experiments for the ACI estimation presented

in the next section. In the fastACI toolbox, tone-plus-noise representations (instead of the default noise-alone representations) are automatically adopted when the experiment is `'modulationACI'`, loading the tone-plus-noise vector $\underline{N_i}$ (here with dimensions 4800-by-1) from the variable `Data_matrix` ($\underline{\underline{N_i}}$, here with 75-by-64 elements).

3.4.2 Statistical estimation

The core principle of the revcorr approach is to assess how the random fluctuations in the stimuli affect the behavioural responses of the participant on a trial-by-trial basis [16]. For this purpose, the revcorr approach relies on a stimulus-response transformation based on a generalised linear model (GLM) to reveal the statistical relationship between the stimuli and the corresponding participant's responses. This transformation results in a T-F matrix of decision weights, the ACI, such that:

$$P_i = \Phi(\underline{N_i}^T \cdot \underline{\mathrm{ACI}} + b) \tag{1}$$

where the predicted value P_i for trial i is a continuous value between 0 and 1 that depends on the noisy vector $\underline{N_i}$ and on the GLM parameters, i.e., the vectorised matrix $\underline{\mathrm{ACI}}$ and the scalar intercept b. The GLM parameters need to be fitted individually based on each participant's data, such that the link function Φ—a logistic function—returns a P_i value close to 1 or 0 when the participant responded $r_i =$ /aba/ (or modulated tone) or $r_i =$ /ada/ (or unmodulated tone), respectively. Each $\underline{\mathrm{ACI}}$ element is interpreted as a perceptual weight of the corresponding T-F point. For this reason, $\underline{\mathrm{ACI}}$ has as many elements as the T-F representations of the targets and its visualisation in matrix form $\underline{\underline{\mathrm{ACI}}}$ can be interpreted in a very straightforward way, with strong positive (excitatory) or negative (inhibitory) perceptual weights being indicated by coloured regions (here red and blue, respectively, see Figures 1 and 2).

Because of the high data dimensionality (up to n = 5504 T-F bins for ABDA22) and the presence of internal noise in the listener's decision, classic maximum-likelihood estimates of the GLM parameters are typically very noisy. It is therefore recommended to use some form of maximum-a-posteriori estimation to constrain the range of GLM solutions by penalising implausible parameter values [17], [18]. Similar to previous studies [5], [18], we adopted an L1-regularisation (also called lasso) using a Gaussian pyramid basis. The Gaussian pyramid is a particular type of basis expansion that expresses the T-F information in a different (multi-scale) system of coordinates. More specifically, we used a 5-level pyramid decomposition applied to each T-F representation, where the T-F matrix was downscaled by a factor of 2 in four successive iterations. For this processing we used the MATLAB function `imresize` but adopting a Gaussian kernel (toolbox parameter `pyramid_script` set to `'imresize'`, see Listing 2). The original T-F representation ("level 0") was discarded from the pyramid (the toolbox parameter `pyramid_shape` was set to 0).

If we define $\underline{\underline{B}}$ as the change-of-basis matrix from the original T-F representation to the Gaussian pyramid space, and $\underline{w}$ as the ACI coordinates in the new space, then Equation 1 can be rewritten as:

$$P_i = \Phi(\underline{N_i}^T \cdot \underline{\underline{B}} \cdot \underline{w} + b) \tag{2}$$

The objective of this change of basis is to formulate the problem in a space where the solutions can be expressed with a limited number of coefficients, meeting the "sparseness" prior imposed by the L1-regularisation. Here, the ACI is described as a linear combination $\underline{w}$ of a limited number of Gaussian-shaped elements of different width (the basis vectors from $\underline{\underline{B}}$). The $\underline{w}$ coordinates were fitted using an L1-penalised logistic regression (default for the toolbox option `glmfct='l1glm'`). The optimal degree of sparsity was chosen to minimise the 10-fold cross-validation deviance of the GLM fitting. Essentially, this method suppresses those weights that do not considerably improve the GLM predictions, while keeping the weights identified as critical for explaining the data. Finally, the vector of weights $\underline{w}$ is displayed in the T-F domain as an ACI, by transforming it back to the original T-F pixel basis:

$$\underline{\mathrm{ACI}} = \underline{\underline{B}} \cdot \underline{w} \tag{3}$$

4 RESULTS

We derived the ACIs for participants SA and SB in all four experiments. The obtained ACIs are shown in the third and fourth columns of Figure 2. To facilitate the later interpretation of results, the spectrograms

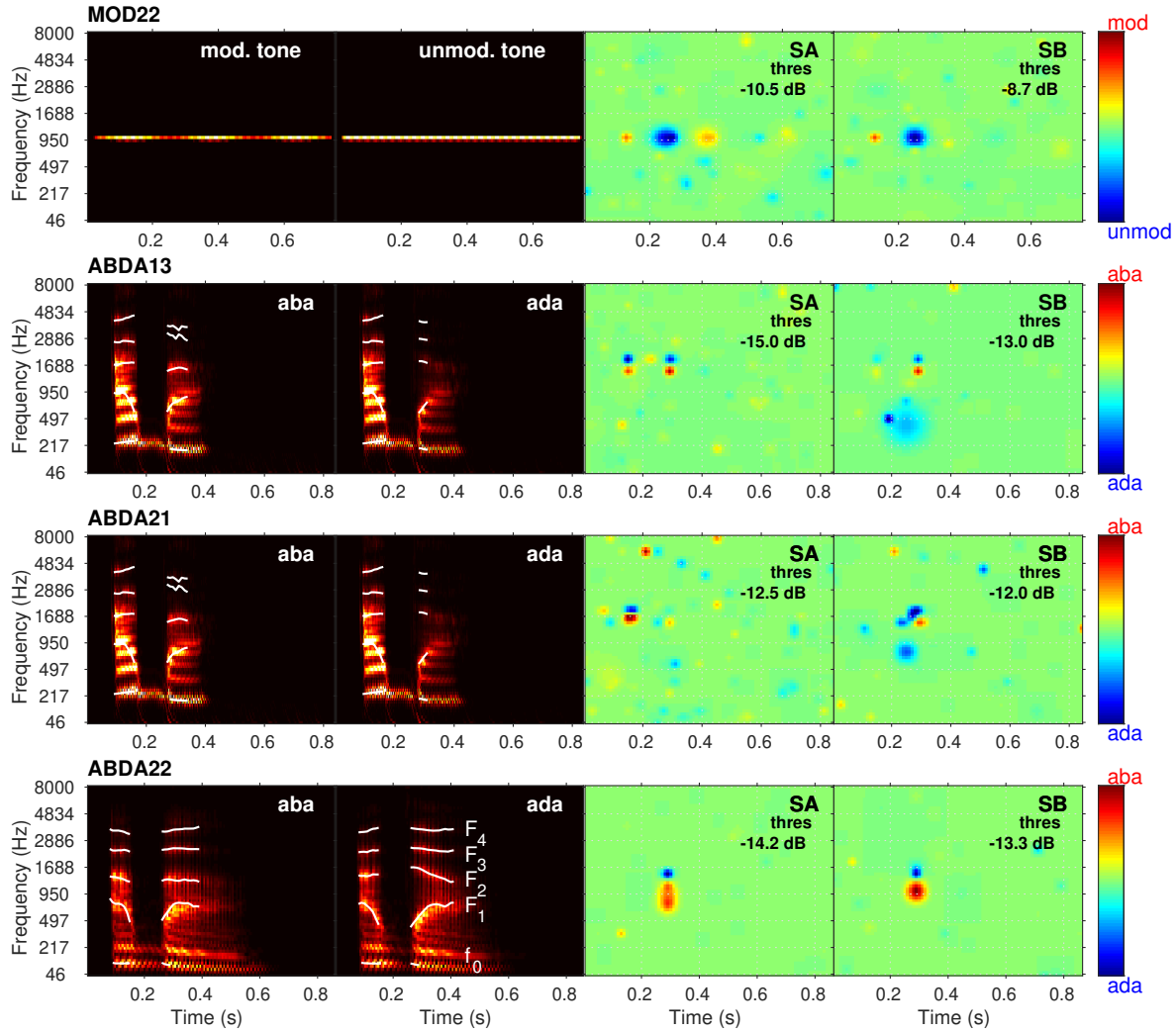

Figure 2. The target sounds and individual ACIs of each experiment are depicted in each of the four rows of panels. The spectrograms of the target sounds (Target 1 and Target 2) are shown in the first two columns, with lighter regions indicating higher signal amplitudes. The last two columns show the corresponding ACIs obtained for participants SA and SB, where positive (red) and negative (blue) weights correspond to T-F regions that biased the participants' responses towards Target 1 or Target 2, respectively. The median thresholds (m_{depth} or SNR) across all N trials for each participant are indicated as insets in the ACI panels.

of the target and reference sounds are shown in the first two panel columns of Figure 2, where darker and lighter regions represent T-F bins with lower and higher amplitudes, respectively. We also superimposed the fundamental frequency (f_0) and the first four formants (F_1–F_4) on the spectrograms of the ABDA experiments. For ease of visualisation we only added the corresponding (f_0 and F_1–F_4) labels in the bottom-most /ada/ spectrogram.

The ACIs show distinct excitatory and inhibitory regions in the T-F space, being this the reason why they are sometimes described as "psychophysical spectro-temporal receptive fields" [19]. In the context of the present experiments, excitatory regions (red positive weights) correspond to T-F configurations of noise energy that result in an increase of the probability that the listener gave the response "/aba/" or "modulated tone". Conversely, inhibitory regions (blue negative weights) correspond to T-F configurations of noise energy resulting in "/ada/" or "unmodulated tone" responses.

4.1 MOD22 experiment

The ACIs for MOD22 are shown in the first row of panels in Figure 2. The excitatory (red) and inhibitory (blue) cues indicate the regions that led participants to answer "modulated tone" and "unmodulated tone," respectively. The ACIs show that noise fluctuations at a frequency around 1000 Hz had a significant contribution in the decision. For both participants, the most dominant cue was located at $t = 0.25$ s and it was of an inhibitory type (blue regions) meaning that the participants answered "unmodulated tone" if they did not hear the first dip in the sinusoidally AM envelope. Other significant but relatively weaker excitatory cues (red regions) were observed at the same frequency (≈ 1000 Hz) but centred at time $t = 0.13$ s and at $t = 0.38$ s (SA) or $t = 0.35$ s (SB), i.e., at the location of the second and third local maxima of the AM envelope which, when perceived, led the participants to answer "modulated tone." All the identified cues are, therefore, oriented horizontally, i.e., along the time dimension. In the discussion section we will focus on the analysis of these cues, and we disregard a number of other less significant cues that were observed primarily for participant SA as, e.g., the (blueish) cue localised at time $t = 0.31$ ms and frequency 300 Hz.

4.2 ABDA experiments

The ACIs for ABDA13, ABDA21, and ABDA22 are shown in the second, third, and fourth row of panels in Figure 2, respectively. The excitatory (red) and inhibitory (blue) cues indicate the regions that led the participants to answer /aba/ and /ada/, respectively. In all panels for both participants there are excitatory and inhibitory cues that are vertically oriented, i.e., along the frequency axis, at time t=0.29 s at around 1500 and 1900 Hz, respectively. At these T-F points is where the consonant-vowel transition is located, including the F_2 formant onset of the /a/ vowel (compare the location of these ACI cues with that of the F_2 traces in the two left-most panels). For participant SB in ABDA13 and ABDA21 (Figure 2, second and third panels), there was an additional inhibitory (blue) cue at $f = 500$ Hz around the F_1 region that was relevant for him to respond /ada/. For these same two experiments, participant SA relied on two vertically oriented cues in the F_2 region of the vowel-consonant transition, at $t = 0.19$ s. In experiment ABDA22, the vertically-oriented cues were only observed at time t=0.29 s, again in the F_2 region. Since the words in this experiment were taken from a male voice, f_0 and the formant frequencies are in a somewhat lower frequency range, with respect to the female voice used in ABDA13 and ABDA21, compare the corresponding spectrograms in Figure 2 (see also Table 2).

5 DISCUSSION

The results from experiments ABDA21, ADBA13 and MOD22 were discussed in detail in our previous studies [2], [4], [5]. In those studies, however, slightly different analyses were performed, all based on experimental ACIs (all experiments) or using simulations (ABDA21, MOD22), with ACIs being estimated using different types of T-F representations and a different GLM algorithm. For comparability purposes, the experimental ACIs presented here were recomputed using the same processing scheme (see Section 3.4). Despite these differences, our findings here are consistent with those of our previous studies, highlighting that the ACIs are only superficially affected by the choice of a particular analysis scheme.

5.1 Acoustic cues

5.1.1 MOD22 experiment

The ACIs for MOD22 from Figure 2 are comparable to the single-band ACIs that Varnet and Lorenzi [4] presented in their Fig. 3**A**. Varnet and Lorenzi calculated the on-frequency ($f = 1000$ Hz) ACI assuming that audio-frequency components far from the tone frequency have a negligible contribution to their modulation-detection task due to the noise masker, i.e., they disregarded the possibility of off-frequency listening. Our 64-band ACI supports that assumption, given that the significant cues are organised horizontally along the carrier frequency of 1000 Hz. Also in line with Varnet and Lorenzi, the dominant cue was found at the location of the first modulation dip at $t = 0.25$ s followed by relatively weaker cues related to the local maxima of the target signal envelope (at $t = 0.13$ s and then between $t = 0.35$ and 0.38 s). Thus, equally-informative acoustic cues in the target modulation (here, the first and second modulation dips) are not weighted equally by the listeners. At the same time, the ACIs reveal an increased perceptual weighting of the first 400 ms—roughly the first half of

the stimulus duration—compared to the later temporal segments. Similar "primacy effects" have been previously reported for other auditory percepts [e.g., 20]. Although the underlying mechanism is still unclear, Varnet and Lorenzi [4] showed, by means of auditory modelling, that such suboptimal weighting could result from transient response characteristics related to auditory modulation processing.

5.1.2 ABDA experiments

The ACIs for experiments ADBA13, ABDA21, and ABDA22 revealed the presence of an acoustic cue in the T-F regions corresponding to the second formant (F_2) transition (between 1500 and 1900 Hz), at the onset of the second syllable (i.e., at 0.29 s for the speech targets used in this experiment). More precisely, all ACIs showed a similar pattern of weights in this region, with a positive cluster below a negative cluster. The weights associated to this cue were very strong except for SA in ABDA21.

This pattern of weights is consistent with the results obtained in previous ACI experiments [2], [6] and confirms the critical role of the F_2 onset as a cue for /b/-/d/ categorisations, as demonstrated using other methods [21]. More specifically, the target sound is more likely to be perceived as /aba/ or /ada/ if the background noise is able to shift the perception of the F_2 formant to a lower frequency (e.g., in the region of the positive weights) or a higher frequency (e.g., in the region of the negative weights), respectively.

A similar configuration of positive and negative weights can be observed in ABDA13 and ABDA21 for participant SA in the F_2 transition at the offset of the first syllable ($t = 0.16$ s, between $f = 1600$ and 1900 Hz). This can be interpreted as evidence for the presence of an anticipatory cue [2]. Because of the overlapping of motor commands during speech production (coarticulation), the identity of the consonant in a VCV word can be inferred from the information in the preceding vowel [22]. Here, SA is relying mainly on this anticipatory cue for solving the task.

The ACIs obtained from the three ABDA experiments suggest that listeners can rely on other speech cues to distinguish between targets, including the F_1 transition near the onset of the second syllable ($t = 0.25$ s, between $f = 400$ and 1000 Hz for participant SB in ABDA13 and ABDA21) and the presence of high frequency energy within the intervocalic interval ($t = 0.21$ s, $f = 6000$ Hz for both participants in ABDA21), matching the T-F position of a potential consonant release burst. The perceptual weights associated to these cues, when present, are weaker and therefore they appear as a secondary source of information for solving the tasks.

5.2 Spectro-temporal cues: Comparison between experiments

Although the ACIs for all experiments revealed a clear pattern of primary acoustic cues, demonstrating that participants are actively extracting information from the noisy stimuli to perform the task. These cues correspond to different characteristics of the targets. In MOD22, positive and negative weights were organised horizontally, and correspond therefore to the detection of a temporal event as a consequence of the (by design) fixed frequency of the tone carrier. On the contrary, in the ABDA experiments, due to the primary role of F_2, the successful completion of the task relied on a fine frequency discrimination (F_2 onset: 1605 Hz for /aba/, 1849 Hz for /ada/ in ABDA13 and ABDA21; 1314 Hz for /aba/, 1467 Hz for /ada/ in ABDA22, see Table 2). This is why the obtained clusters of weights around F_2 are arranged vertically in Figure 2.

For ABDA experiments, where several conditions were evaluated, there are a number of interesting observations that can be extracted from the comparison of ACIs for different background noises as used in ABDA13 (white noise) and ABDA21 (SSN), both using the same speech sounds from a female speaker, or from the comparison between ACIs using sounds recorded from a female or male speaker, as in ABDA13 and ABDA22, respectively. We present these observations next.

5.3 ABDA13 and ABDA21: Comparison between white-noise and SSN backgrounds

When comparing the ACIs from the second (ABDA13) and third panel (ABDA21) of Figure 2, we can observe that the relevant T-F cues were overall located in similar time and frequency regions in both conditions. In other words, the participants did not seem to modify considerably their listening strategy across noise conditions. Instead, only the relative importance of the different cues was changed. For instance, while participant SA used F_2 cues in both syllables in the white-noise condition (second panel), he provided a stronger weight to the first syllable and only a weak weight to the second one in the SSN condition (third panel). Participant SB

used primarily the information in the second syllable, more specifically, around the F_2 region to choose /aba/ or /ada/, and around F_1 to choose /ada/ (blue region). For this participant, the cues in the first syllable were weak in the white-noise condition but even weaker in the SSN condition.

It is well known that not all acoustic cues are equally robust to a given type of background noise [23]–[25]. For instance, the high-frequency burst cue in a /t/ consonant is more easily perceived in a SSN than in a white-noise condition, because white-noise maskers have more energy in high-frequencies [24]. These differences across background conditions raise the question of how the extraction of one cue or another depends on the listening situation. Very little is known about this phenomenon, as it is very difficult to study experimentally.

In another example, Serniclaes and Arrouas [26] used artificially-edited stimuli to show that listeners weight acoustic cues differently when phonemes are perceived in a silent or in a noisy condition, with low-frequency cues being very vulnerable to noise. The authors interpreted this result as a shift in the listening strategy towards the most reliable cues as the noise increased. More recently, in a study using natural speech stimuli (/aba/–/ada/ or /alda/–/alga/–/arda/–/arga/ words) for which ACIs were obtained for original and degraded versions of the stimuli, Varnet *et al.* [27] suggested that the weighting of cues was not only determined by their robustness to noise but was also likely to be related to how subjectively relevant the cues were for the specific participant. Such top-down effect was supported by the observation of anticipatory cues and low-frequency F_1 cues, which were observed even when they were not really present in the acoustic properties of their stimuli.

The present study supports this hypothesis, as we also observed a similar anticipatory effect (participant SA in ABDA13 and ABDA21) and a low-frequency F_1 cue (participant SB in ABDA13 and ABDA21), while the spectral distribution of SSN and white-noise maskers did not seem to affect considerably the primary (and "robust") F_2 cues.

5.4 ABDA13 and ABDA22: Comparison between speech utterances

When comparing the ACIs from the second (ABDA13) and fourth panel (ABDA22) of Figure 2, we can observe that the ACIs for ABDA22, where sounds from the male speaker were used, do not show a salient cue in the first syllable when compared to the results obtained using the female speaker sounds. So, SA and SB did not benefit of any anticipatory cue when the male speech sounds were used. To explain this difference in the T-F cues, we found that the /aba/ and /ada/ sounds from the female speaker of ABDA13 had a very intense first syllable compared to the second syllable ($\Delta L = 5.5$ and 7.7 dB for /aba/ and /ada/ respectively, see Table 2) whereas in the male speaker sounds of ABDA22, the intensity of the syllables was more balanced with a level difference ΔL of 3.1 dB or slightly less. Hence, when the SNR adaptively reaches the point where the primary cue for the task (F_2 onset, second syllable) is just audible, the secondary cue (F_2 offset, first syllable) is much more likely to be audible—2.4 or 4.6 dB more intense in /aba/ or /ada/—in ABDA13 than in ABDA22. This is consistent with previous studies of phoneme-in-noise perception, that demonstrate that some utterances are more robust to noise than others, depending on the intensity of the primary cue and the presence of conflicting cues [23], [28].

5.5 Discussion summary

In this section we discussed observations that we could draw from the obtained ACIs and we contextualised them with respect to the existing literature. We illustrated the power of the ACI method for (1) the precise characterisation of individual T-F cues for a modulation-detection task and for a set of three /aba/-/ada/ experiments (Section 5.1), (2) identifying the nature of the relevant perceptual cues either temporal or spectral (Section 5.2), (3) identifying subtle changes in the listening strategy such as those by different types of noise maskers (Section 5.3) by adopting different speech targets (Section 5.4).

6 CONCLUSIONS

In this illustrative contribution of the revcorr method, we showed the results of two participants in four listening experiments (ABDA13, ABDA21, ABDA22, and MOD22), that were summarised in Table 1 and schematised in Figure 1. The experiments used different types of noisy listening conditions (white noise or speech-shaped noise), combined with an /aba/-/ada/ categorisation task or a modulation-detection task. A total of four audi-

tory classification images (ACIs) for each participant were obtained, revealing the specific set of time-frequency (T-F) cues on which the participants relied—their listening strategy—during the course of the corresponding experiment. The obtained ACIs were similar to those obtained previously [2], [5], [8], demonstrating the reliability of this novel psychophysical paradigm. Although our discussion was purely based on the information that could extract from the ACIs presented in Figure 2, a much more elaborate characterisation of the assessed listening strategies can be performed, including estimates of prediction power of the ACIs and how is this actually affected by the specific acoustic properties of the noises [8], [9]. As one of the disadvantages of the revcorr method is the large number of trials that need to be collected to assess reliable ACIs (N between 3000 and 5000 trials in the evaluated experiments), our current efforts are posed into the optimisation, on the one hand of the revcorr implementation, and on the other of the ACI estimation. To allow a transparent and straightforward optimisation of these two aspects of the method, we implemented the whole framework in an in-house MATLAB toolbox that we named fastACI. The fastACI toolbox does not only contain all our working framework but it was also implemented modularly to allow future stage-by-stage extensions.

ACKNOWLEDGEMENTS

This study was funded by the ANR grants "fastACI" (Grant No. ANR-20-CE28-0004) and "FrontCog" (Grant No. ANR-17-EURE-0017).

A APPENDIX: Study replication using the fastACI toolbox

The four experiments presented in the current study can be replicated using our in-house fastACI toolbox for MATLAB [3]. We start this appendix by indicating how to replicate the experiments (Section A.1) and how to reproduce or replicate the ACI estimation (Section A.2). We finish this appendix by briefly commenting on the installation and current development status of the fastACI toolbox (Section A.3). First-time toolbox users should start in Section A.3.

A.1 Replication of the experiments

Assuming that you have a computer equipped with MATLAB and you have downloaded and initialised the toolbox (see Section A.3), a fastACI experiment can be run using the commands in Listing 1, preceded by one of the following input options:

```
%% ABDA13 experiment:
experiment = 'speechACI_varnet2013';
Condition = 'white'; % default condition
```

```
%%% ABDA21 experiment:
experiment = 'speechACI_varnet2013';
Condition = 'SSN'; % non-default, needs to be specified
```

```
%% ABDA22 experiment:
experiment = 'speechACI_Logatome-abda-S43M';
Condition = 'white'; % default condition
```

```
%% MOD22 experiment
experiment = 'modulationACI';
Condition = 'white'; % default condition
```

Listing 1: Reproducing a fastACI experiment, for an arbitrary participant 'S01'.

```
Subject_ID = 'S01';
fastACI_experiment(experiment,Subject_ID,Condition);
```

After the completion of all N trials, the toolbox will have generated a time-stamped `*.mat` or "savegame" file. For instance, for participant `"S01"` in ABDA22, the savegame file will have a name of the type: `savegame_2022_03_23_13_30_S01_speechACI_Logatome-abda-S43M_white.mat`.

A.2 Reproduction or replication of the ACI estimation

Listing 2: MATLAB code required to generate the ACI for a result file named "savegame_2022_03_23_13_30_S01_speechACI_Logatome-abda-S43M_white.mat."

```
glmfct = 'l1glm'; % L1 (lasso) GLM function
TF_type = 'gammatone'; % Type of T-F conversion
flags_in = { 'trialtype_analysis','total', ...
             'N_folds', 10, ...
             'no_permutation', ...
             'no_bias', ...
             'plot', ... % or set to 'no_plot'
             'pyramid_script','imresize', ...
             'pyramid_shape',0 };
fname_results='savegame_2022_03_23_13_30_S01_speechACI_Logatome-abda-S43M_white.mat';
[ACI,cfg_ACI,results,Data_matrix]=fastACI_getACI(fname_results,TF_type,glmfct,flags_in{:});
```

A.3 Installation and development status of the fastACI toolbox

To install the fastACI toolbox, after its download from https://github.com/aosses-tue/fastACI [3], you need to initialise the toolbox by running the script `startup_fastACI`. At least one third-party package is needed, the AMT toolbox (https://amtoolbox.org/ in its version 1.0 or more recent). After initialisation you should be able to run all the codes shown in this appendix.

The fastACI toolbox has been inspired by features from two toolboxes: AMT [29] and AFC [30]. AFC toolbox users will notice the similarity of our toolbox for the implementation of experiments: All experiments consist of a set of scripts. The three main types are: `*_set.m` containing fixed experimental parameters, `*_cfg.m` containing general configuration settings, `*_user.m` containing the trial definition.

From the AMT toolbox, we adopted a similar way to reproduce figures from the literature. The reproduction of figures from our previous studies are contained in `publ_*.m` files (located in the toolbox folder ../Publications/), which are similar to the `exp_*.m` scripts in AMT. For instance, to reproduce Figs. 1A from [5] (ACI for participant SA in ABDA21, similar to the ACI in Figure 2, but using "old ACI settings") can be obtained by typing `publ_osses2021c_DAGA_figs('fig1a')`. In fastACI, however, this and other scripts may require the manual download of data or the preparation of data before being able to use the `publ_*_figs.m` file.

A.4 Citing the fastACI toolbox

The fastACI toolbox can be cited as a "project" (all versions) or pointing to a specific version of the toolbox. To cite all versions (it points to the latest release) use:

Osses & Varnet (2022). fastACI toolbox: the MATLAB toolbox for investigating auditory perception using reverse correlation. doi:10.5281/zenodo.5500138. GitHub: github.com/aosses-tue/fastACI

REFERENCES

[1] A. Ahumada and J. Lovell, "Stimulus features in signal detection", *J. Acoust. Soc. Am.*, vol. 49, pp. 1751–1756, 1971. DOI: 10.1121/1.1912577.

[2] L. Varnet, K. Knoblauch, F. Meunier, and M. Hoen, "Using auditory classification images for the identification of fine acoustic cues used in speech perception", *Front. Hum. Neurosci.*, vol. 7, pp. 1–12, 2013. DOI: 10.3389/fnhum.2013.00865.

[3] A. Osses and L. Varnet, *fastACI toolbox: the MATLAB toolbox for investigating auditory perception using reverse correlation (v1.1)*, 2022. DOI: 10.5281/zenodo.6530154.

[4] L. Varnet and C. Lorenzi, "Probing temporal modulation detection in white noise using intrinsic envelope fluctuations: A reverse correlation study", *J. Acoust. Soc. Am.*, vol. 151, 2022. DOI: 10.1121/10.0009629.

[5] A. Osses and L. Varnet, "Consonant-in-noise discrimination using an auditory model with different speech-based decision devices", in *DAGA*, 2021, pp. 298–301. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03345050.

[6] L. Varnet, K. Knoblauch, W. Serniclaes, F. Meunier, and M. Hoen, "A psychophysical imaging method evidencing auditory cue extraction during speech perception: A group analysis of auditory classification images", *PLoS one*, vol. 10, no. 3, pp. 1–23, 2015. DOI: 10.1371/journal.pone.0118009.

[7] L. Varnet, C. Lorenzi, and A. Osses, "Probing amplitude-modulation detection and phoneme categorization with auditory reverse correlation", in *Congrès Français d'Acoustique*, Marseille, 2022. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03641680.

[8] A. Osses and L. Varnet, "Auditory reverse correlation on a phoneme-discrimination task: Assessing the effect of different types of background noise", in *ARO mid-winter meeting*, 2022. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03553443v1.

[9] A. Osses and L. Varnet, *A microscopic investigation of the effect of random envelope fluctuations on phoneme-in-noise perception*, OSF Preregistration, 2022. [Online]. Available: osf.io/ya6v7.

[10] H. Levitt, "Transformed up-down methods in psychoacoustics", *J. Acoust. Soc. Am.*, vol. 49, no. 2, pp. 467–477, 1971.

[11] C. Kaernbach, "Simple adaptive testing with the weighted up-down method", *Percept. Psychophys.*, vol. 49, pp. 227–229, 1991. DOI: 10.3758/BF03214307.

[12] B. Meyer, T. Jürgens, T. Wesker, T. Brand, and B. Kollmeier, "Human phoneme recognition depending on speech-intrinsic variability", *J. Acoust. Soc. Am.*, vol. 128, pp. 3126–3141, 2010. DOI: 10.1121/1.3493450.

[13] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data", *Hear. Res.*, vol. 47, pp. 103–138, 1990. DOI: 10.1016/0378-5955(90)90170-T.

[14] A. Osses and A. Kohlrausch, "Perceptual similarity between piano notes: Simulations with a template-based perception model", *J. Acoust. Soc. Am.*, vol. 149, pp. 3534–3552, 2021. DOI: 10.1121/10.0004818.

[15] A. Osses, L. Varnet, L. Carney, T. Dau, I. Bruce, S. Verhulst, and P. Majdak, "A comparative study of eight human auditory models of monaural processing", *Acta Acust.*, vol. 6, p. 17, 2022. DOI: 10.1051/aacus/2022008.

[16] R. Murray, "Classification images: A review", *J. Vis.*, vol. 11, pp. 1–25, 2011. DOI: 10.1167/11.5.2.

[17] K. Knoblauch and L. Maloney, "Classification images", in *Modeling psychophysical data in R*, Springer, 2012, ch. 6, pp. 167–194. DOI: 10.1007/978-1-4614-4475-6_6.

[18] P. Mineault, S. Barthelmé, and C. Pack, "Improved classification images with sparse priors in a smooth basis", *J. Vis.*, vol. 9, pp. 1–24, 2009. DOI: 10.1167/9.10.17.

[19] D. Shub and V. Richards, "Psychophysical spectro-temporal receptive fields in an auditory task", *Hear. Res.*, vol. 251, pp. 1–9, 2009. DOI: 10.1016/j.heares.2009.02.007.

[20] E. Ponsot, P. Susini, G. Saint Pierre, and S. Meunier, "Temporal loudness weights for sounds with increasing and decreasing intensity profiles", *J. Acoust. Soc. Am.*, vol. 134, EL321–6, 2013. DOI: 10.1121/1.4819184.

[21] A. Liberman, P. Delattre, and F. Cooper, "The role of selected stimulus-variables in the perception of the unvoiced stop consonants", *Am. J. Psychol.*, vol. 65, pp. 497–516, DOI: 10.2307/1418032.

[22] P. Warren and W. Marslen-Wilson, "Continuous uptake of acoustic cues in spoken word recognition", *Percept. Psychophys.*, vol. 41, pp. 262–275, 1987. DOI: 10.3758/BF03208224.

[23] S. Phatak, A. Lovitt, and J. Allen, "Consonant confusions in white noise", *J. Acoust. Soc. Am.*, vol. 124, pp. 1220–1233, 2008. DOI: 10.1121/1.2913251.

[24] M. Régnier and J. Allen, "A method to identify noise-robust perceptual features: Application for consonant /t/", *J. Acoust. Soc. Am.*, vol. 123, pp. 2801–2814, 2008. DOI: 10.1121/1.2897915.

[25] L. Varnet, J. Meyer, M. Hoen, and F. Meunier, "Phoneme resistance during speech-in-noise comprehension", in *Interspeech*, Portland, USA, 2012, pp. 599–602. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00932285.

[26] Y. Serniclaes W.; Arrouas, "Perception des traits phonétiques dans le bruit", *Verbum*, pp. 131–144, 2 1995.

[27] L. Varnet, F. Meunier, and M. Hoen, "Speech reductions cause a de-weighting of secondary acoustic cues.", in *Interspeech*, 2016, pp. 645–649. DOI: 10.21437/Interspeech.2016-343.

[28] J. Zaar and T. Dau, "Sources of variability in consonant perception of normal-hearing listeners", *J. Acoust. Soc. Am.*, vol. 138, pp. 1253–1267, 2015. DOI: 10.1121/1.4928142.

[29] P. Majdak, C. Hollomey, and R. Baumgartner, "AMT 1.x: A toolbox for reproducible research in auditory modeling", *Acta Acust.*, vol. 6, p. 19, 2022. DOI: 10.1051/aacus/2022011.

[30] S. Ewert, "AFC - A modular framework for running psychoacoustic experiments and computational perception models", in *Proceedings of the International Conference on Acoustics AIA-DAGA*, 2013, pp. 1326–1329.