forum acusticum 2026
8-12 September · Graz, Austria

ASSESSING THE EFFICACY OF MARKOV CHAIN MONTE CARLO WITH PEOPLE TO ESTIMATE THE MENTAL REPRESENTATION OF VOWELS

Mauro Manucci[1], Ladislas Nalborczyk[2], Léo Varnet[1]

[1] *Laboratoire des systèmes perceptifs, École normale supérieure, PSL University, CNRS, Paris, France*
[2] *Aix-Marseille University, CNRS, LPL, France*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Vowel recognition relies on specific physical properties of sound, known as "acoustic cues", that listeners use to identify phonemes. The first two formant frequency values (F1 and F2) have been repeatedly shown to be the primary acoustic cues for vowel categorisation. However, identifying the exact distribution underlying their mental representation remains a difficult task. Current methods, such as reverse correlation, are very time intensive as they typically require thousands of trials. In the present study, we tested whether the method Markov Chain Monte Carlo with people (MCMCp) with adaptive Metropolis-Hastings sampling could provide a more efficient alternative to probe the mental representation of vowels. In a perceptual two-interval, two-alternative forced choice task with synthetic vowels, we showed that vowel-specific regions in F1-F2 space were more precisely estimated using the MCMCp approach than with the reverse correlation method, given the same number of trials. These regions were also congruent with those reported in the production literature. Furthermore, these results do not critically depend on the number of features considered. The use of the MCMCp may therefore prove highly useful for experimental phonetics.

Keywords: *markov chain monte carlo with people, reverse correlation, vowel perception, klatt synthesizer, mental representation*

1. Introduction

The identification of perceptual primitives (or "acoustic cues") underlying phoneme perception is a central and long-standing question in psycholinguistics. Vowel sounds, for instance, are defined by a range of acoustic features, including duration, fundamental frequency, and formant frequencies (F1, F2, F3, and F4). A fundamental challenge is determining which of these acoustic features the auditory system relies on for vowel recognition.

Historically, acoustic phonetics has provided foundational insights into this question by examining the distribution of these features across productions of the same vowels and across different vowels. Phonetictians have mapped the acoustic space of vowels, offering a baseline for understanding their perceptual organisation [1], [2].

F1 and F2, which are critical for distinguishing vowel sounds, define a 2-dimensional space where vowels occupy distinct regions. For example, the vowel /u/ is characterised by a high, back tongue position, corresponding to lower F1 and F2 formant frequencies, while /i/ corresponds to low F1 and high F2 (see Figure 1). It may be therefore assumed that listeners hold a prototype representation in F1-F2 space for every vowel, a single idealised exemplar that is contrasted with incoming sound stimuli in order to make a decision on their identity. However, production data represented in Figure 1 are only indirectly informative about perceptual phenomena. Furthermore, the F1-F2 space presents a simplified picture: as previously noted, vowels are characterised by multiple acoustic features, not all of which may be equally relevant. Hence, given the high dimensionality of the sound stimulus, there is a need for new techniques that efficiently probe such distributions.

Over the last decades, reverse correlation has emerged as a powerful experimental tool for identifying acoustic cues [4], [5], [6]. In a typical reverse-correlation experiment, participants perform a categorisation task based on stimuli where random fluctuations have been introduced by the experimenter. By correlating these fluctuations with participants' responses, researchers derive a "kernel" that estimates the influ-

***Corresponding author**: leo.varnet@cnrs.fr.*

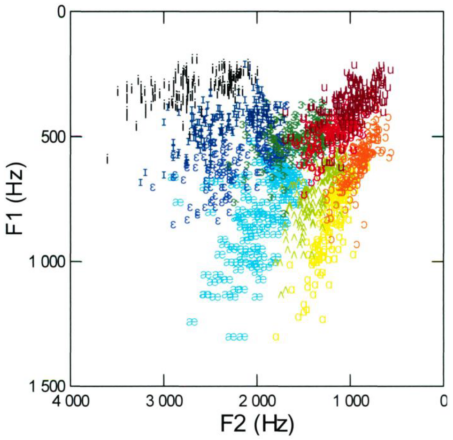

Figure 1: Dispersion diagram F1/F2 for ten monosyllables produced by men, women and children. The figure was adapted from [3], Figure 3, and the production data gathered by [2]. The target vowels relevant to the present study are /ɛ/ (blue), /[u/ (dark red) and /ɑ/ (yellow).

ence of each acoustic feature on categorisation. In 2013 [7], this approach was applied within a vowel perception paradigm where listeners were presented with noise-only stimuli and were asked to indicate whenever they heard [a] or [iː]. Consistent with production studies [1], [2], the resulting kernels, obtained by correlating noise spectrograms to the detection of the target vowel, revealed critical regions near the locations of F1, F2, and possibly F3.

However, a major drawback of reverse correlation is its time-intensive nature. For instance, in [7], participants had to complete one hour of testing, while studies aiming for higher precision required more than 2.5 hours per participant [5]. The objective of the present project is to improve the efficiency of this estimation process, in order to reduce experimental duration. Two complementary options are considered to achieve this goal. The first one is dimensionality reduction. In any statistical model, the number of trials required for obtaining a reliable estimation is directly tied to the number of predictors. Therefore, the duration of the experiment can be reduced by considering only a well-chosen subset of dimensions (the six acoustic features mentioned above). We aim to streamline the experimental process by implementing a variation of the "dimensional noise" technique [8]. With this approach, the sound stimulus parameters are directly varied using a speech synthesizer, rather than applying white noise to a reference.

The second solution for improving efficiency is replacing the current "offline" paradigm, where all noise stimuli are generated prior to the testing, with an adaptive approach. In this framework, the next stimulus to be presented depends on the data already collected. A promising candidate for this adaptive algorithm is Markov Chain Monte Carlo with people (MCMCp).

The MCMCp method has been previously implemented to model abstract mental representations in the visual domain with great success, including animal shapes and facial affect categories [9], [10]. In essence, the MCMCp aims at exploring the parameter space more efficiently through an iterative process, generating trajectories that converge to a stationary distribution over time, with centre corresponding to a prototypical representation of the target. To the best of our knowledge, there has been only one attempt at applying this approach to the exploration of phoneme categories [11]. However, it was limited in scope, as it only varied F1 and F2 over a discretized stimulus space, with only one target vowel assigned to no more than six participants, and it lacked any comparison with standard approaches. The present study will therefore aim to probe the mental representation of vowels with the MCMCp method more systematically and contrast its efficacy with reverse correlation.

2. Materials and methods

2.1. Participants and materials

Participants were recruited on Prolific. Inclusion criteria were being an English native speaker, being over 18 and under 50 years of age, and not having any kind of speech or hearing impairment. Unbeknownst to them, each participant was randomly assigned to one of three experiments, reverse correlation, MCMCp-reduced or MCMCp-full, and one of three target vowels (see below). Their participation in any one of the experiments excluded them from partaking in the others. They received a monetary compensation after completion. 21 participants were recruited for each MCMCp experiment and 9 participants for the reverse correlation experiment. For the MCMCp-full and MCMCp-reduced experiments, 12 and 10 participants were excluded from further analyses as they did not meet convergence criteria (see section 3.2).

2.2. Procedure

The experiment was coded in PsychoPy (PsychoJS, ver. 2024.2.4) and hosted online on Pavlovia. Stimuli were generated with a TypeScript/JavaScript implementation of the Klatt synthesiser [12], [13] on a trial-to-trial basis. The sound stimuli had 50 ms long ascending and descending cosine ramps.

The listeners were randomly assigned to one of three vowels: [u], [ɑ] or [ɛ], and completed a two-interval, two-alternative forced choice (2I-2AFC) task. At the beginning of the experiment, they were prompted with a text that instructed them to choose the stimulus, the first or the second, that best approximated a target vowel. Depending on the condition, the target vowel was described as "u" as in "boot", "e" as in "bet", or "a" as in "ask".

The session consisted of three blocks with 200 trials each, yielding a total of 600 choices for every experiment. Across all included and excluded participants, the median session lengths were of 35, 40, and 37 minutes for the reverse correlation, MCMCp-reduced, and MCMCp-full experiments, respectively. At the

end of the experiment, participants completed a survey assessing their dialect, bilingualism or additional languages spoken, naturalness of the stimuli, as well as how close the final sounds were to the target vowel.

2.3. Algorithm

2.3.1. Parameters and boundaries

Of all parameters available in the Klatt synthesizer, only a subset were manipulated during the experiment. In the reverse correlation and MCMCp-reduced experiments, only 3 dimensions of interest (F1, F2 and duration) were varied. For the MCMCp-full experiment, six parameters were varied, $f0$, F1-F4 and stimulus duration. All other parameters pertaining to vowel synthesis were kept constant.

The frequency boundaries of the formants of interest were defined based on the minimum and maximum values for each formant across all vowels tested in production data [1], [14], and reports on optimal Klatt synthesizer parameters [15]. That is, each dimension of interest had a different range of possible values, and these ranges were constant across all experiments and target vowel conditions (Table 1). Furthermore, a constraint on fundamental frequency and formant order (i.e., $f0 < F1 < F2$, etc.) was ensured throughout. As a result, in practice, our method violates the MCMC assumption of ergodicity, since the whole parameter space cannot not be freely explored without breaking the aforementioned constraints. Nevertheless, this ostensible violation of ergodicity is necessary to make sure that the stimuli are perceptually valid.

Table 1: Parameter ranges.

Parameter	min	max
$f0$	80 Hz	180 Hz
F1	140 Hz	810 Hz
F2	600 Hz	2300 Hz
F3	1900 Hz	2900 Hz
F4	2700 Hz	4130 Hz
duration	700 ms	1000 ms

At initialisation, a vowel was generated with values for the dimensions of interest, independently sampled from a uniform distribution within each parameter's range. Then, a second stimulus ("proposal stimulus") with different parameter values was generated based on the rules of the specific experimental condition (see paragraphs Reverse correlation and MCMCp below). Both stimuli were presented in random order and the participant's task was to decide which one of two consecutively presented sounds best approximated their target vowel. At the next trial, the same procedure was iterated with the chosen stimulus and a new proposal stimulus. All three experiments had repeated samples, as a participant could choose the same proposal for several trials in a row.

2.3.2. Reverse correlation

In the reverse correlation condition, the proposal stimulus was simply drawn from a uniform distribution, independent of the previous responses of the participant. This amounts to randomly sampling stimuli from the parameter space.

2.3.3. MCMCp

The MCMCp experiments followed an implementation of the Metropolis-Hastings (MH) algorithm with adaptive step size. The proposal parameters of interest were each drawn from a Gaussian distribution centred on the current stimulus. A total of six, one hundred trials-long, chains were completed by each participant (see Figure 3 for traceplot examples).

Furthermore, the standard deviation of the Gaussian distribution was made adaptive in order to probe the parameter space more efficiently. It followed a simple update rule: $\epsilon_{t+1} \leftarrow \epsilon_t \cdot e^{\gamma \cdot \delta}$, where the constant $\gamma = 0.65$ is the adaptation rate, and δ is the difference between the target acceptance rate (0.75) and the current acceptance rate. Acceptance rate was computed from the past history of a 22 trial-long sliding window. Target acceptance rate, adaptation rate and sliding window length were defined based on pilot data, such as on the number of trials it took participants to reach their target vowel, and on whether they would get stuck hearing stimuli that did not sound like their target vowel. In addition, for the MCMCp experiments, chains were interleaved; that is, unbeknownst to the participant, they were alternatively choosing from two simultaneous 100 trials-long chains per block. This was done to prevent participants from noticing the repetition of stimuli.

3. Results

3.1. Single participant examples

Figure 2 shows an example of the trajectories of two of the six chains in the F1-F2 space from a single participant with target vowel [ɑ]. After a few trials, the chains appear to converge, then oscillate around a stable value. Exploration then continues more locally, as the step size is adaptively decreased and the chains stay in the same region of the parameter space.

All six chains from the same participant are represented as a traceplot in Figure 3, where parameters F1 and stimulus duration are plotted as a function of trial number. It can be observed that while all six F1 chains seem to converge to a higher region of frequencies, duration does not seem to follow any kind of pattern.

3.2. Convergence

Upon visual inspection, it becomes clear from the F1 traceplot (Figure 3) that there comes a point after which the participant entered a stable region of the state space. Determining the exact trial at which this phase ends has been the topic of extensive discussion in the field of the MCMC [16]. Trials from this transition (or "burn-in") phase are typically discarded before analysing the properties of the convergence region. We hence aimed to test whether, over time, participants converged to a stationary distribution.

Assuming that F1 is the most relevant parameter for determining the end of the transient phase, that point was here defined as the trial at which the ratio between the cross-chain F1 standard deviation

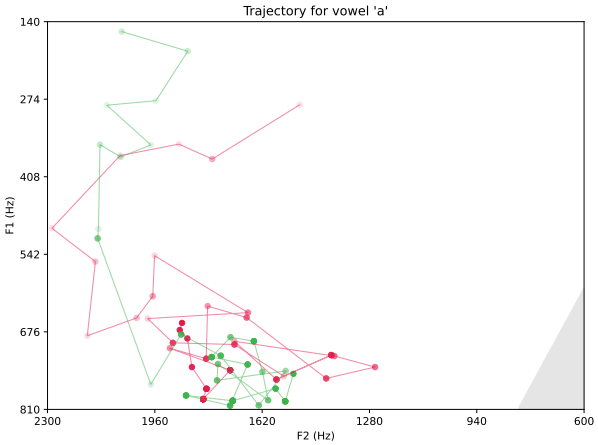

Figure 2: Trajectory example for a single participant. Evolution of two chains, of the six total, in F1-F2 space; each was assigned a different colour. Greater colour saturation reflects later trial progression. Note that rejected proposals are not shown in the figure, and that some points have repeated samples. The shaded area represents the forbidden region where F1 > F2. Axis boundaries directly correspond to the state space boundaries.

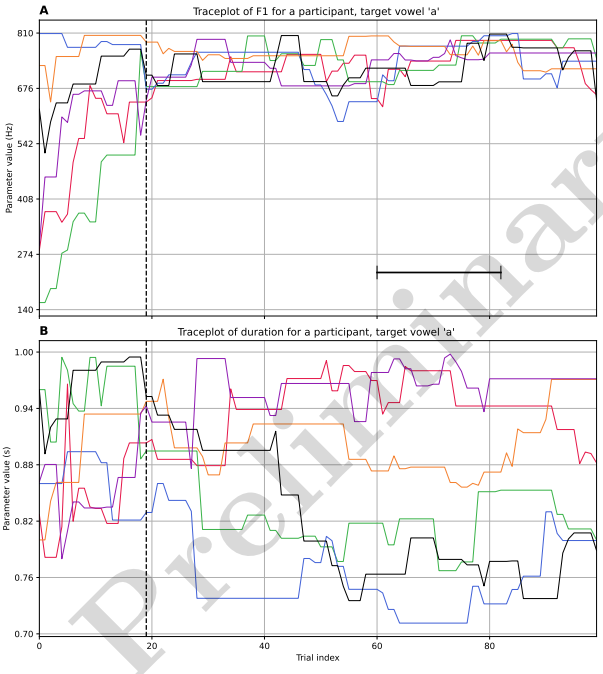

Figure 3: Traceplot examples from the same participant displayed in Figure 2. Traceplots showing the evolution of chosen F1 (A) and chosen duration (B) across time for all six chains. From visual inspection, F1 converges to a stable region after a burn-in phase, whereas duration does not. The dashed lines mark the end of the transient phase based on the defined convergence heuristic (see section 3.2) and the floating segment in A represents the trial window length chosen for step size adaptability.

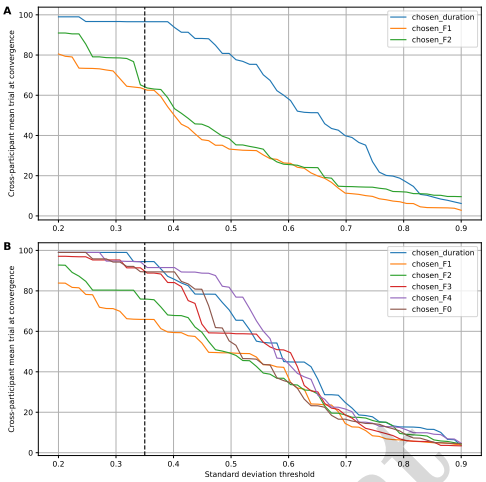

Figure 4: Average trial at convergence for the MCMCp-reduced (A) and MCMCp-full (B) experiments, as a function of the standard deviation threshold free parameter. F1 and F2 tend to converge earlier than the rest of dimensions of interest for lower threshold values. Dashed line indicates the chosen threshold value (0.35).

and the standard deviation of a uniform distribution with the range of the F1 dimension, remained under a certain threshold for at least 5 consecutive trials. A motivation for choosing this heuristic was that the cross-chain standard deviation from all participants at onset is independent of target vowel and it is proportional to the range of the dimension of interest allowing for direct comparisons. Figure 4 shows the cross-participant average trial in which each parameter converged under the previous metric, as a function of the standard deviation convergence threshold. As expected, this relationship is negative and monotonic. It can be seen that, for threshold values lower than 0.4, F1 and F2 reach convergence earlier than any other parameter. Conversely, the function becomes difficult to interpret, with respect to parameter convergence, for threshold values higher than 0.45 since, after that point, most parameters acquire on average a similar time-to-convergence value. As a result, a threshold of 0.35 was selected to define the end of the transient phase. Trials before the estimated burn-in endpoint were discarded, and participants who did not reach convergence before 75 trials were excluded from the analysis. This heuristic kept 9 participants of the 21 recruited listeners from the MCMCp-full experiment, and kept 11 of the 21 total from the MCMCp-reduced experiment. No performance exclusion criterion was applied to the reverse correlation experiment.

3.3. Probed F1-F2 distributions

After convergence, pooled data representing the distribution in F1-F2 space from chosen synthetic vowels can be seen on Figure 5, representing the total average heatmap results across converged participants for each vowel group and experiment. Additionally, the first and third quartiles of the distribution are reported in

Table 2. Note that for the reverse correlation case, no convergence method was applied, so all participants and trials were used. Regions of greater density in heatmaps were mostly consistent across conditions, although they were much noisier for reverse correlation. The regions corresponding to the different vowels generally matched those identified in previous production studies (Figure 1).

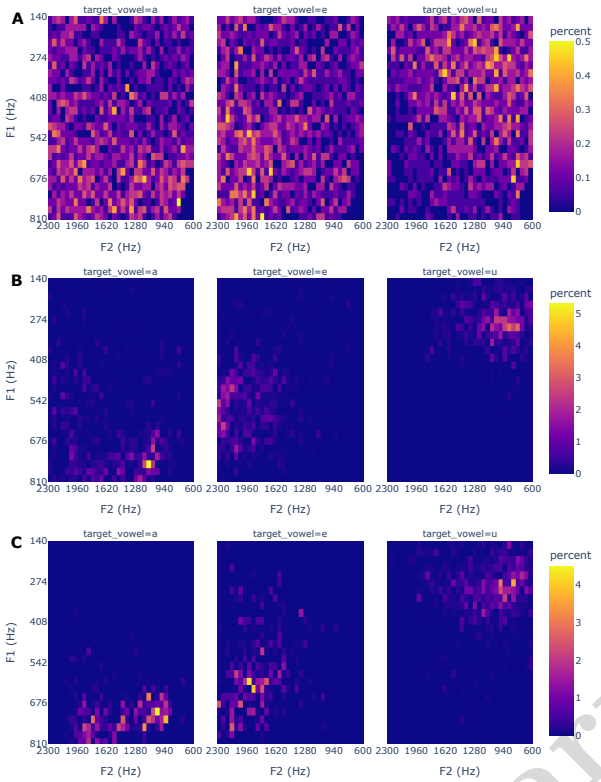

Figure 5: Heatmaps of the counts of selected stimuli pooled across participants in F1-F2 space, normalised (%) by the total selection counts for each vowel (columns [ɑ], [ɛ] and [u], in order) and experiment (rows). (A) Reverse correlation, (B) MCMCp-reduced, (C) MCMCp-full. Bin size is of 25 Hz by 50 Hz for the F1 and F2 dimensions, respectively. For the MCMCp figures, only trials after convergence were utilised.

Table 2: Quartiles Q_1-Q_3 of the F1 and F2 distributions for each condition and target vowel (in Hz).

Experiment		[ɑ]	[ɛ]	[u]
MCMC-full	F1	690-755	558-670	251-336
	F2	1057-1745	1735-2032	844-1284
MCMC-red	F1	626-763	484-635	260-317
	F2	1124-1895	1794-2186	833-1153
RevCorr	F1	365-682	338-658	270-547
	F2	1080-1945	1194-1985	960-1665

3.4. Effect of dialect

The listeners' dialects were quite varied, with participants reporting to belong to different North American, European and African dialect groups. Most counts were American - Midwest, British - East Midlands, and South African, with five counts each, from all three experiments combined. Visual inspection revealed no apparent consistent bias elicited by dialect groupings.

4. Discussion

In the present study, we aimed at probing the mental representation of vowels using the MCMCp method, and contrast its efficiency with that of the reverse correlation method. Both F1 and F2 seemed to have converged efficiently and to regions congruent with previous production data (see Figure 1), both in absolute frequency values, as well as for the relative positionings in the F1-F2 space.

Although we would expect the reverse correlation figure to approximate the MCMCp figures as the total number of trials increases, we found that the region identified in the F1-F2 space with the MCMCp approach was considerably more localised than with the reverse correlation method. This suggests that the MCMCp requires less trials than the reverse correlation approach to model the mental representation of vowels to the same degree of precision.

For both MCMCp experiments, and for low standard deviation thresholds, F1 and F2 reach convergence earlier on average than any other parameter (Figure 4). This is in line with previous literature that showed that humans rely primarily on F1 and F2 when categorising vowels. However, the gap between F1 and F2 seems to vanish earlier for the MCMCp-reduced experiment than the MCMCp-full experiment.

Interestingly, although high density centres identified in this study (Figure 5) are close to those reported in the vowel production literature (Figure 1), they seem to be slightly deviated towards the closest boundary. For example, for the case of [u], it can be seen that it has a considerably lower first formant mode than the typical array of F1 values measured for that vowel reported from production data [1]. This was likely not due to an issue pertaining to the specifically chosen JavaScript package implementation of the Klatt, as corroborated with Praat and another Klatt package implementation on Python, which behaved in the same manner. Potentially, the observed pattern could have been a result of a preference for "overarticulated" vowels over more naturally produced ones in the 2I-2AFC task, which may shift F1 and F2 closer to their nearest boundary. Remarkably, the count density of F2 was much more spread than for F1 (Figure 5), both in absolute frequency as well as relative to the dimension range, suggesting that listeners rely primarily on F1 for vowel identification.

The limited ecological validity of the synthetic vowels might be another possibility for the observed deviation from production data. The advantage of the Klatt synthesiser is its easiness of use, at the cost of potentially less natural sounds. Under the current method, vowels lacked a time-dependent variation of the formants, such as those described in [14]. Moreover, participants

gave generally low "naturalness" scores, with a mean rating of 2.33 ± 1.22 SD (on a 1-5 scale) across the three experiments.

Despite the well-established methods that exist to assess chain convergence in the MCMC, there is a lack of a standardised convergence metric in the MCMCp literature. One of the most widely used statistics is $\hat{R}$, which is proportional to the ratio of the within-chain variance over the between-chain variance. According to the $\hat{R}$ criterion, a between-chain variance that closely approximated the agent's internal noise should yield a value not that much greater than 1.0. Conversely, trajectories that do not converge hold a much greater cross-chain variance than the internal (within-chain) variance, yielding $\hat{R} \gg 1$. Not all studies in the MCMCp literature explicitly cite the use of the $\hat{R}$ statistic, with some relying on visual inspection ("thick pen" technique) and chain-crossing counts [9], [10]. Notably, the present convergence method is more objective than the thick pen technique and, according to analyses not reported in this paper, it was also more robust than the chain crossing counts heuristic. Although our current convergence method was well adjusted with visual inspection of the traceplots, it does not consider cases of high internal noise (either derived from the participant or intrinsic to a target vowel with a sparse formant distribution) like the $\hat{R}$ statistic does. These caveats are soon to be addressed with simulation data.

The experiment also aimed to test differences between a lower dimensional (MCMCp-reduced) and a higher dimensional (MCMCp-full) space. Visual inspection revealed no apparent differences between either method in F1-F2 space distributions by using either of these methods (Figure 5). This suggests that our method works well independent of the number of parameters considered. This becomes especially important when considering that traditional synthetic speech experiments are limited to the exploration of one or two parameters at a time. On the contrary, MCMCp allows probing a large number of parameters, which is more efficient, less hypothesis-driven, and also more ecological. However, F2 converges, on average, slightly later in the MCMCp-full experiment (Figure 4), hinting at the possibility that the addition of more dimensions could, conversely, make participants less reliant on the lower formants and thereby delay convergence.

In short, in the present study we were able to probe the distribution of the dimensions of interest for three vowels using the MCMCp and reverse correlation methods via the presentation of synthetic vowels. The MCMCp identified a more localised region of the F1-F2 space for the same number of trials as the reverse correlation method, while providing a close replication of the typical formant values reported for the target vowels considered. Further research should focus on convening on a standardised method for assessing convergence and trial discard after burn-in. These caveats will be soon tackled with a follow-up simulation study.

References

[1] J. Hillenbrand, L. A. Getty, K. Wheeler, and M. J. Clark, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical Society of America*, vol. 95, pp. 2875–2875, May 1994.

[2] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *The Journal of the Acoustical Society of America*, vol. 24, pp. 175–184, Mar. 1952.

[3] C. Sigouin, "Caractéristiques acoustiques des voyelles fermées tendues , relâchées et allongées en français québécois," M.S. thesis, 2013.

[4] L. Varnet, K. Knoblauch, F. Meunier, and M. Hoen, "Using auditory classification images for the identification of fine acoustic cues used in speech perception," *Frontiers in Human Neuroscience*, vol. 7, 2013.

[5] G. Carranante, C. Cany, P. Farri, M. Giavazzi, and L. Varnet, "Mapping the spectrotemporal regions influencing perception of french stop consonants in noise," *Scientific Reports*, vol. 14, Nov. 2024.

[6] A. Osses and L. Varnet, "A microscopic investigation of the effect of random envelope fluctuations on phoneme-in-noise perception," *The Journal of the Acoustical Society of America*, vol. 155, pp. 1469–1485, Jan. 2024.

[7] W. O. Brimijoin, M. A. Akeroyd, E. Tilbury, and B. Porr, "The internal representation of vowel spectra investigated using behavioral response-triggered averaging," *The Journal of the Acoustical Society of America*, vol. 133, EL118–EL122, Jan. 2013.

[8] J. J. Burred, E. Ponsot, L. Goupil, M. Liuni, and J.-J. Aucouturier, "Cleese: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition," *PLOS ONE*, vol. 14, P. Loui, Ed., Apr. 2019.

[9] J. B. Martin, T. L. Griffiths, and A. N. Sanborn, "Testing the efficiency of markov chain monte carlo with people using facial affect categories," *Cognitive Science*, vol. 36, pp. 150–162, Oct. 2011.

[10] A. N. Sanborn, T. L. Griffiths, and R. M. Shiffrin, "Uncovering mental representations with markov chain monte carlo," *Cognitive Psychology*, vol. 60, pp. 63–106, Mar. 2010.

[11] J. Pooley, "Exploring phonetic category structure with markov chain monte carlo," M.S. thesis, 2008.

[12] C. d'Heureuse, *Klatt-syn: Klatt formant synthesizer*, GitHub, 2019. Accessed: Feb. 16, 2026. [Online]. Available: `https://github.com/chdh/klatt-syn/tree/master`

[13] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *The Journal of the Acoustical Society of America*, vol. 67, pp. 971–995, Mar. 1980.

[14] P. Ladefoged and K. Johnson, *A Course in Phonetics*. Wadsworth Cengage Learning, 2010.

[15] J. Coleman and A. Slater, "Estimation of parameters for the klatt synthesizer from a speech database," *Data-Driven Techniques in Speech Synthesis*, pp. 215–238, 2001.

[16] B. Lambert, *A Student's Guide to Bayesian Statistics*. SAGE, Apr. 2018.

5. Acknowledgments

This study was funded by the ANR-DFG grant "DRhyADS" (ANR-22-FRAL-0003) to Léo Varnet and Alessandro Tavano, and the EUR Frontiers in Cognition ANR-17-EURE-0017.