PNAS
www.pnas.org

Supplementary Information for

Speakers are able to categorize vowels based on tongue somatosensation

Jean-François Patri, David J. Ostry, Julien Diard, Jean-Luc Schwartz, Pamela Trudeau-Fisette, Christophe Savariaux and Pascal Perrier

Pascal Perrier.
E-mail: Pascal.Perrier@grenoble-inp.fr

This PDF file includes:

Supplementary text
Figs. S1 to S7
References for SI reference citations

Supporting Information Text

Recordings and data processing

Acoustical signals were recorded at 44,100 Hz using a C-1000S AKG microphone. Tongue and lip movements were recorded using an Electromagnetic Articulography system (EMA Wave-NDI) at 200 Hz and then low-pass filtered at 30 Hz for data analysis. Four EMA sensors were glued on the tongue in the mid-sagittal plane: one on the tongue tip, two on the tongue body and one on the tongue dorsum (see left panel of Fig 1 in the main text). Two additional EMA sensors were placed on the upper and lower lip. Subjects were provided with real time visual feedback of the lip sensor positions and were asked to keep them as stationary as possible during the experiment. Sensors were glued on with PeriAcryl 90HV, a high viscosity Cyanoacrylate Oral Adhesive safe to use on oral tissues. A 6-DOF NDI reference sensor was placed on the nasion and was used to extract the relative position of the sensors. The left panel in Fig 1 in the main text shows the location of each of these 7 EMA sensors.

The tongue postures retained for data analysis were automatically identified based on the acoustical signal. To do so, the start and end of each voiced or whispered sound were detected based on an acoustical energy threshold. Then, within this period, a 1-second interval of greatest articulatory stability was selected (minimal average standard deviation of the positions of the four tongue sensors). The reached tongue posture for each trial was chosen as the tongue configuration that was the closest to the median value within the 1-second time window of best stability.

We discarded data from trials in both the production task and tongue positioning task for which tongue position variability was greater than two standard deviations from the average variability of all subjects in the production task. This corresponded to a maximum variation of 3 mm during the 1-second interval. In total 19 out of 400 trials were discarded from the production task and 10 out of 720 trials were discarded from the tongue positioning task (see below Fig S4 for violin plots representing the tongue position variability for each trial and the discarded trials in both the production and tongue positioning tasks).

Data analysis

Entropy of auditory answers. The entropy $H_s(t)$ of auditory answers, associated with tongue posture t of subject s (Fig 4 in main text) was computed from the probability $P_s^A(v \mid t)$ that tongue posture t of subject s would be attributed to vowel $v \in \{/e/, /\varepsilon/, /a/\}$. We estimated these probabilities as the ratio of the number n_V of auditory answers attributed to vowel v, over the total number of auditory answers n_{Tot} for this tongue posture:

$$P_s^A(v \mid t) = \frac{n_V}{n_{Tot}}. \quad [1]$$

Each reached tongue posture was labeled by each of the eight auditory subjects, which resulted in eight auditory answers per tongue posture, so $n_{Tot} = 8$. The entropies $H_s(t)$ are then computed as:

$$H_s(t) = -\sum_v P_s^A(v \mid t) \ln P_s^A(v \mid t). \quad [2]$$

Values of entropy of auditory answers for each whisper. Data points of Fig 4 in the main text correspond to the entropy of the auditory answers given by eight listeners to the whispered speech. Data points are distributed along ten discrete entropy levels which correspond to the ten possible patterns of distribution of the eight auditory answers given to each whisper of /e/ or /ɛ/ or /a/, without considering the vowel labels: [8 0 0], [7 1 0], [6 2 0], [6 1 1], [5 3 0], [5 2 1], [4 4 0], [4 2 2], [4 3 1], [3 3 2]. The first item [8 0 0] corresponds to the configuration with the lowest entropy value (zero entropy), where all eight auditory subjects agreed about the phonetic labeling of the heard sound; the last item [3 3 2] corresponds to the configuration with the highest entropy value (entropy of 1.08), where three listeners agreed on a phonetic label, three others agreed on an other phonetic label, and the remaining label was selected by two listeners.

Clustering scores - Silhouette analysis. The silhouette analysis provides a quantitative evaluation of the consistency and the separability of clusters by calculating how similar each observation is to the other observations in the cluster it is assigned to, as compared to how similar it is to the observations in the other clusters (1). In this study tongue postures are assigned to three possible vowel clusters /a/, /e/ and /ɛ/ on the basis of somatosensory or auditory labels. The quality of clustering is evaluated by computing a silhouette score $s(t)$ for each tongue posture t as:

$$s(t) = \frac{b(t) - a(t)}{\max\{a(t), b(t)\}}, \quad [3]$$

with $a(t)$ being the average distance between t and all other data within the cluster it is assigned to, and $b(t)$ being the smallest average distance of t to all data in all other clusters. Silhouette scores range from -1 to 1, with positive values corresponding to observations that are well within their assigned cluster, values close to 0 corresponding to observations that are at the borderline with neighboring clusters, and negative values corresponding to observations that are closer to a neighboring cluster than to their assigned cluster, suggesting that they may be assigned to the wrong cluster. Silhouette values are commonly visualized in the form of horizontally displayed bar graphs (presented below in Fig S5 and Fig S6), and their average value provides an overall measure of clustering consistency. Bars in Fig 5 in the main text present these average scores for each subject.

The Bayesian classifier. The classifier computes, for each somatosensory subject s, the probability that tongue posture t is categorized as vowel v. This computation is performed as Bayesian inference from the joint probability distribution $P(v\ t\ s)$ that we decompose as:

$$P(v\ t\ s) = P(s)P(v)P(t \mid v\ s). \quad [4]$$

We consider $P(s)$ and $P(v)$ as uniform probability distributions and we specify $P(t \mid v\ s)$ from the statistics of tongue postures performed during the production task as follows. We modeled $P(t \mid v\ s)$ as Gaussian probability distributions $\mathcal{N}(t; \mu_s^v, \sigma_s^v)$ describing the token-to-token variability of tongue postures produced by each subject s during the production of each vowel v. We estimate the mean and covariance parameters (μ_s^v, σ_s^v) from the statistics of tongue postures measured for the 10 repetitions of each vowel during the production task. In order to reduce the dimensionality of tongue postures (8D for 4 EMA sensors in 2-dimensional sagittal plane) we describe the tongue postures of each particular subject in the 2-dimensional space defined by the first two principal components obtained from a Principal Component Analysis of all vowel tongue postures produced by the subject during the production task. The proportion of variance accounted for by these first two principal components ranged between 93% and 99% (average 96%) across subjects.

Performing Bayesian inference from the decomposition of Eq. 4 leads to

$$P(v \mid t\ s) = \frac{P(t \mid v\ s)}{\sum_v P(t \mid v\ s)}. \quad [5]$$

Construction of identification curves. The identification curves present the proportion of labels that are attributed to a given vowel, along the continuum of tongue postures defined for each somatosensory subject. We represent the continuum of tongue postures along a single dimension corresponding to the ordering of the nine target tongue postures $TTP_s = \{T^k; k \in [1:9]\}_s$ of each subject s from /e/ to /a/. To construct these curves we began by subdividing the set of all reached tongue postures $RTP_s = \{T^j; j \in [1:90]\}_s$, performed by subject s, into nine subsets $C_s^k = \{T^l\}_s^k$ corresponding to the nine target tongue postures. These subsets were formed by associating each reached tongue posture in the RTP_s set with its closest target tongue posture in the TTP_s set, in terms of Euclidean distance. We then computed, for each subset C_s^k, the fraction of labels that are attributed to each vowel v. In the case of the somatosensory identification curves, this computation only considered the answers provided by the corresponding somatosensory subject. In the case of the auditory identification curves though, this computation includes all answers provided by all auditory subjects. In the case of the identification curves predicted by the Bayesian Classifier for each somatosensory subject, the proportions of labels attributed to vowel v within each subset C_s^k were computed by averaging the probabilities $P_s(v \mid t)$ assigned by the Bayesian Classifier to each of the tongue postures t within the considered subset C_s^k.

Could the whispering task provide auditory cues for vowel identification?

The design of the present experiment required that subjects whisper before making a judgment of vowel identity based on felt tongue position. For the subsequent somatosensory vowel identification task, it is crucial that the subjects could not base their judgements on auditory information from their own whispers. We had taken a number of precautions in order to minimize the likelihood that this could happen, by adding masking noise an by displaying a sound level-meter that helped subjects keep the intensity of their whispers low. We evaluated the acoustic power of the recorded whispers and found them to be on average between 35 and 45dB-SPL, depending on the subject, whereas the masking noise was 80dB-SPL. In our previous work, the auditory perception of French vowels in noise was found not to be possible at this signal to noise level (2).

We also considered the possible contribution of bone conduction to the auditory perception of the whispers. Since whispered speech does not involve vibrations of the vocal folds, its bone conducted contribution to the signal at the cochlea is limited (3). In order to confirm this for the present data, we conducted the following quantitative test. Based on Howell's & Powell's (1984) (3) measurements of the bone vibration transfer function during speech, we estimated a filter simulating bone acoustic conduction for the articulations that we considered in this experiment. Using this filter we estimated that, for the recorded whispers, a reduction of acoustic power at the level of the cochlea would be on average between -15 and -20dB compared to the whispered acoustic power at the lips.

The results of both analyses are consistent with subjects' reports at the end of experiment that they could not hear themselves whisper.

Could the subjects use speech motor control cues to identify vowels during the whispering task?

The tongue position during the whispering phase of the positioning task was not totally stable in all trials in spite of instructions given to participants. We assessed whether the observed tongue movements might reflect particular strategies used by the subjects to identify vowels.

First, we hypothesized that, if movements during the whispering phase were performed intentionally by subjects in order to facilitate somatosensory vowel identification, subjects who performed well in the somatosensory identification task would present more articulatory variability, both in terms of the amount of variation and in terms of the number of variable trials. Fig S1A,C,E present three measures quantifying for each subject the variability of tongue position during the whispering phase. Panels A and C show the average and the maximum of variation within each trial. Panel E provides a global measure of variability computed as the proportion of "non-stable trials" among all trials performed by each subject (see the Methods

Section below for the characterization of "non-stable trials"). These three measures of variability indicate that subjects who performed well in the somatosensory identification task (S3, S6, S7, S8) did not produce trials that were more variable than those of the other subjects, either in terms of the amount of variability or of the number of variable trials. Furthermore these three measures indicate that for all subjects most trials were stable. In particular, Panel A indicates that for all subjects except S5, more than 75% of trials had an average deviation smaller than half the distance separating two neighboring target postures (postures represented by solid and dotted lines on Panel B of Fig 1 of the main text). Panel C further shows that for all subjects except S5, more than 50% of the trials have a maximum variability that is smaller than the average distance between two neighboring target postures. This is further supported by Panel E which shows that less than 50% of trials are "non-stable". Taken together, these analyses indicate that tongue movements during whispering did not systematically occur in all trials (less than 50% of trials) and were not specific to the subjects who better performed in the somatosensory identification task.

Second, if the variability in tongue posture during the whispering phase was related to a strategy that was intended to identify the closest prototypical vowel posture, we would expect that less variable trials would be those in which the tongue posture reached at the beginning of the whispering phase (called "initial tongue posture" henceforth) was closer to one of the three vowel target postures (solid lines on Panel B of Fig 1 of the main text). Conversely, the most variable trials should be those for which the initial tongue posture was closer to one of the six intermediate target postures (dashed lines on Panel B of Fig 1 of the main text). Panels B and D show the average and maximal variations of tongue posture split according to whether the initial vowel posture was closer to one of the vowel target postures or closer to a "non-vowel" intermediate target posture, for the four subjects who performed well in the somatosensory identification task (S3, S6, S7, S8). No significant difference was found between the two classes of trials, either for the average variation (Fig S1B) or for the maximum variation (Fig S1D); unpaired unequal variance t-test, $p>0.05$ for all subjects in both measures of deviation). This is further supported by Fig S1F showing that for all subjects except S6 the proportion of "non-stable trials" that deviate from an initial tongue posture that is close to a vowel-target is not significantly different from the proportion of trials that deviate from an initial tongue posture that is closer to an intermediate target posture (two-tailed two-proportions z-test, $p>0.05$ for all subjects, $p<0.001$ for S6). Note that the specificity of subject S6 in this regard is due to a particularly low proportion of variable trials with an initial posture that is closer to a vowel target tongue posture, and not to a particularly large proportion of variable trials with an initial posture that are closer to intermediate tongue postures (not significantly greater in S6 than in S3 or S7 (two proportion z-test, $p > 0.05$).

Third, we assessed whether "non-stable trials" tended to converge towards vowel target postures more often than towards intermediate targets postures. If tongue posture variation during whispering is not intentionally driven towards vowel target postures but results from a random process, the probability of converging towards a vowel target posture is 1/3 since there are three vowel-targets among the total of nine targets. Fig S1G shows that for all subjects except S6 the proportion of "non-stable trials" that converge towards a vowel target posture is not significantly greater than the 1/3 probability of chance (one-tailed one-proportion z-test, $p > 0.05$ for all subjects, $p<0.001$ for S6).

Finally, visual inspection of variation in tongue position during the whispering phase reveals that a significant proportion of "non-stable trials" are characterized by a progressive drift of the tongue. This drift does not converge towards vowel target postures more than towards intermediate target postures, as demonstrated by our analyses above. Rather tongue position appears to drift for each subject in the same direction across trials, independently from relation between the initial posture in the whispering phase and the closest vowel target posture. To confirm this observation we extracted the direction of drift of each trial using linear regression. Wilcoxon signed rank test confirmed that the distribution of drift direction across trials was significantly biased towards a given direction ($p < 0.001$) for each subject. This further indicates that tongue movements during whispering were not associated to a particular strategy of subjects developed to facilitate vowel identification, but rather reflect the intrinsic difficulty of the task.

In sum, for subjects S3, S7 and S8 all the analyses support the hypothesis that tongue movements during whispering did not provide motor cues for the subsequent somatosensory identification task. In regard to subject S6, the articulatory variability seems to differ whether or not the initial posture in the whispering phase was or was not a vowel posture. However, the other results showing in particular that for this subject the average and maximal articulatory variations and the proportion of "non-stable" trials with an initial posture that is closer to an intermediate target did not differ from those of the subjects S3 and S7 suggests that our conclusion should apply to subject S6 as well.

Methods. Tongue posture variations shown in Fig S1A-D were computed along the main direction of the PCA carried out on the nine target tongue postures (as defined above in Section "Assessment of the uniform coverage of the task workspace defined by the target tongue postures") and with respect to the reached tongue posture that we considered to be representative of the trial (as defined in the main text in subsection "The whispering task" of "Material and Methods"). We categorized a trial as a "non-stable trial" when the initial tongue posture during the whispering phase did not match the reached tongue posture that we extracted during the most stable phase of the whisper. The initial tongue posture during the whispering phase was extracted as the median posture across the first 200ms of this phase. The initial and reached postures were considered to match if the closest target tongue posture was the same for both of them.

Assessment of the uniform coverage of the task workspace defined by the target tongue postures

In order to propose a relevant characterization of the task workspace defined for each subject by the nine target tongue postures, we carried out, for each subject separately, a Principal Component Analysis on these postures. We found that for all subjects

but S2, tongue target postures were very well described by a single dimension: the variance accounted for by the first principal component was more than 96% (89% for S2).

In this context we then assessed whether such a single dimensional description also held true for the reached tongue postures. To do so we carried out a second Principal Component Analysis again for each subject separately on the target tongue postures. Fig 3B in the main text shows the proportion of variance explained by the first two principal components resulting from this second PCA. There are two clusters, the first one including subjects S3, S6, S7 and S8, the second one including S1, S2, S4 and S5. This clustering is confirmed by a hierarchical clustering analysis (see Fig S2). In the first group, the reached postures are well represented by a single dimension, with the first principal component accounting for more than 88% of the variance, and the second principal component accounting for less than 10%; for the second group this effect is less strong, with the first principal component accounting for an average of 78%. An unpaired unequal variance t-test indicated that the proportion of variance accounted for by the first principal component differed significantly between these two groups of subjects ($t = 6.91$, $p = 0.001$). A measure of the similarity (cosine similarity) between the first principal components, computed on the target postures and on the reached tongue postures, showed that for all subjects but S2 the two directions differed by less than 18 degrees.

Taken together these two results indicate that subjects S3, S6, S7 and S8 reached tongue postures that were in good agreement with the variation of the target tongue postures within the articulatory range of vowels /e/, /ɛ/ and /a/, and essentially varied along a single dimension in the corresponding articulatory workspace. For the other subjects the reached tongue postures did not match well with the task workspace defined by their target tongue postures.

We next tested whether or not there was uniform coverage of the task workspace by the reached tongue postures. To do so we estimated the densities of the reached tongue postures of each subject along the dimension defined by the first principal component of the target tongue postures using MATLAB® ksdensity function. In each panel of Fig 3A of the main text the right side shows the resulting distribution over the range of the nine target tongue postures. We used the Kolmogorov-Smirnov test to evaluate the hypothesis that the obtained distribution of reached tongue postures was compatible with a uniform distribution. For all subjects apart from S1 and S5 there were no statistically reliable differences from uniformity ($p > 0.05$).

Combining these results, we concluded that among the 8 subjects tested only subjects S3, S6, S7 and S8 were able to achieve the tongue positioning task in a way that is compatible with a further assessment of the somatosensory identification of vowels: they were the only subjects that had reached tongue postures which essentially varied along a single dimension that was compatible with the main dimension of variation of the target tongue postures, and had reached postures that uniformly covered the full /e, ɛ, a/ range.

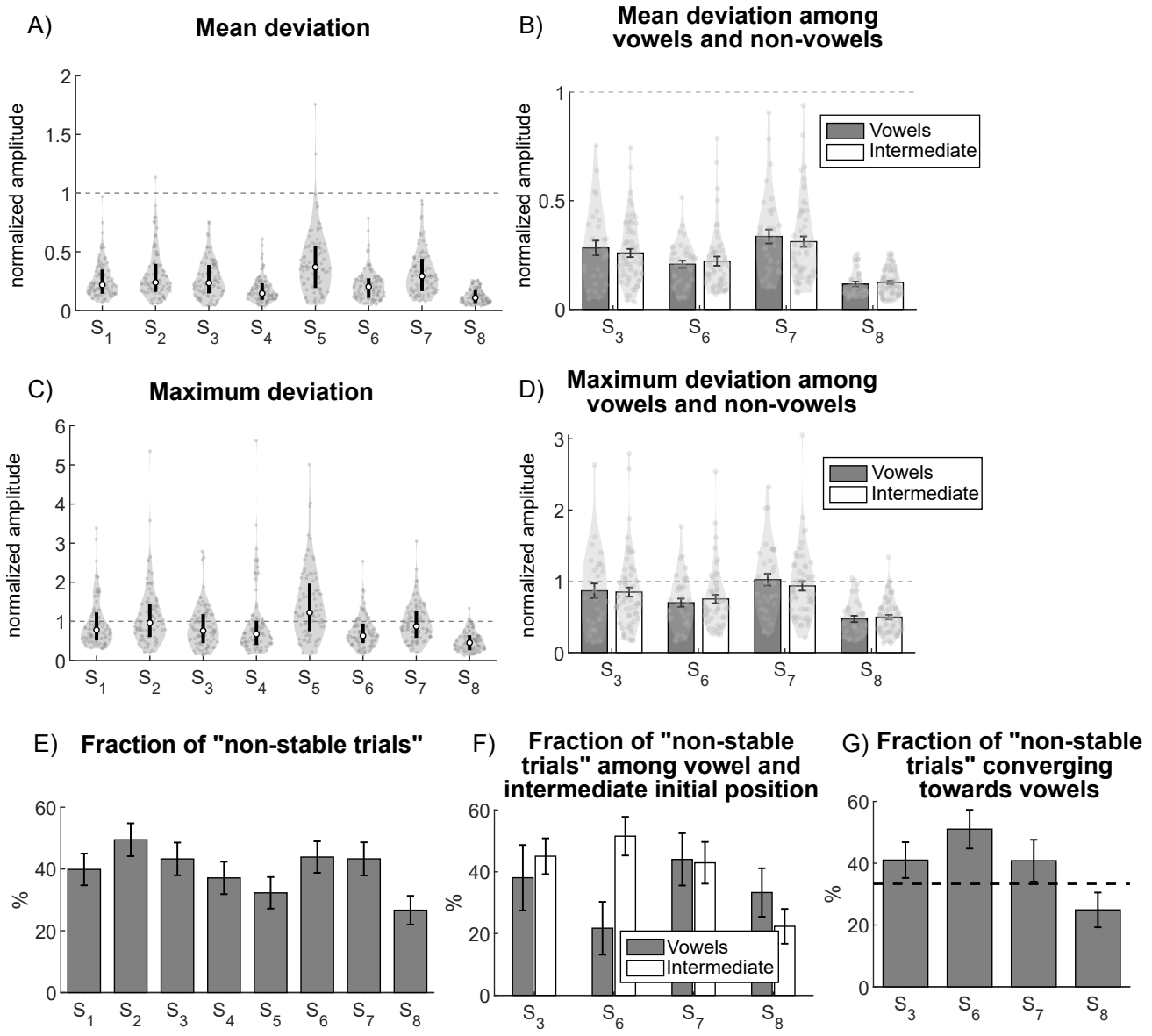

Fig. S1. Analyses of tongue posture variation during the whispering phase. Panels A, B, C and D display measures of the magnitude of tongue posture variation normalized for each subject by the average distance between two neighboring target postures. Panels A,C: Variation of tongue posture across trials for each subject, quantified by its average (panels A) and its maximum (panels C) with respect to the reached tongue posture extracted during the most stable part of the trial. White circles indicate the median of the distribution of the tongue posture for each subject and the black vertical bars indicate the 25% and 75% percentiles. Panels B,D: Comparison of the variability of trials in which the initial tongue posture was closest to a vowel target posture and of trials in which the initial tongue posture was closest to an intermediate target posture (panels B:average variation, panel D: maximum variation). Panel E: Proportion of "non-stable trials" among all trials for each subject. Panel F: Proportion of "non-stable trials" among the trials in which the initial tongue posture corresponded to a vowel target posture and among the trials in which the initial tongue posture corresponded to an intermediate target posture. Panel G: Proportion of "non-stable trials" that converge towards a vowel target posture. Error bars in panels indicate standard error of the mean, computed from 2000 bootstrap samples.

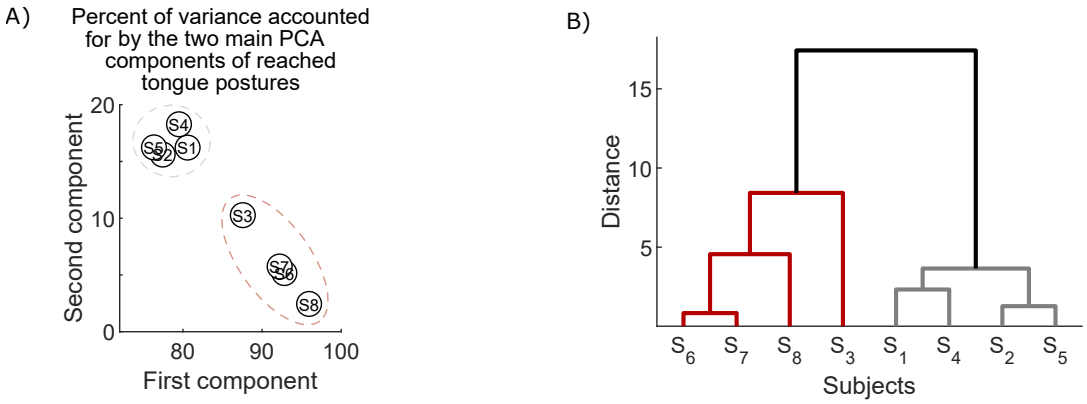

Fig. S2. Hierarchical clustering of somatosensory subjects based on the percent of variance accounted for by the first and second component resulting from the Principal Component Analysis performed, for each subject separately, on the set of reached tongue postures. Panel A replicates panel B of Fig 3 of the main text. The separation into two clusters as indicated by the elliptical regions (drawn by hand) in panel A is supported by the dendrogram in panel B, constructed based on the euclidean distance between subjects in the plane displayed in panel A.

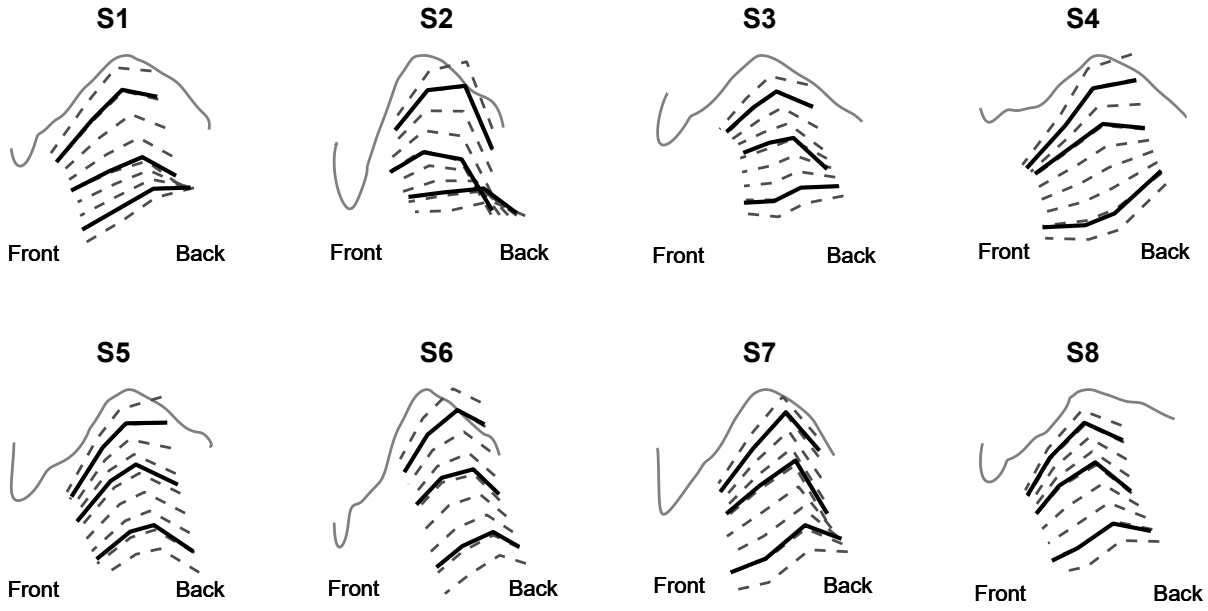

Fig. S3. Set of target tongue postures used in the tongue positioning task for each somatosensory subject (sagittal view). Plain lines represent vowel target tongue postures, dashed lines represent intermediate target postures. The upper grey trace represents the subject's palate. Note that there is an overlap between the tongue contours and the palatal trace for some subjects (S2, S4, S5 and S6). This is due to inaccuracies in the palatal trace (obtained by moving a coil attached to a finger along the palate in the midsagittal plane of the subject). A small shift from the exact midsagittal plane can easily occur, with greater visible effects for subjects having arched palates than for those having flat palates. This has no consequences at all for our analyses, since the palatal contour is only used for the visual representation of tongue postures.

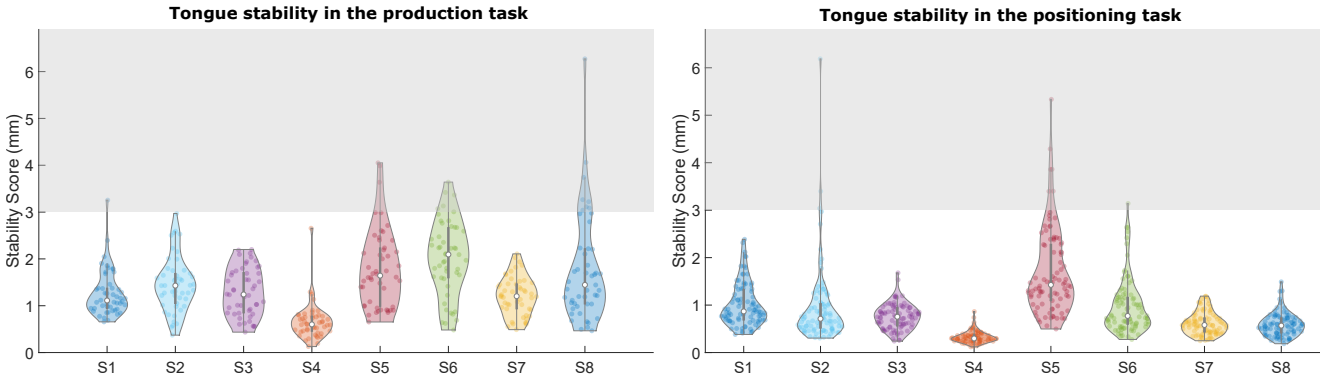

Fig. S4. Violin plots for each subject showing the variability of tongue position during the 1-second-best-stability time window retained for each trial in our analyses, during the production task (panel A) and the positioning task (panel B). Variability is computed as the square root of the variance computed in the 8-dimensional space of the coordinates of the four tongue coils in the sagittal plane. Trials exceeding two standard deviations of all trials of all subjects in the production task were considered as outliers and were removed from our analyses (data in the grey areas).

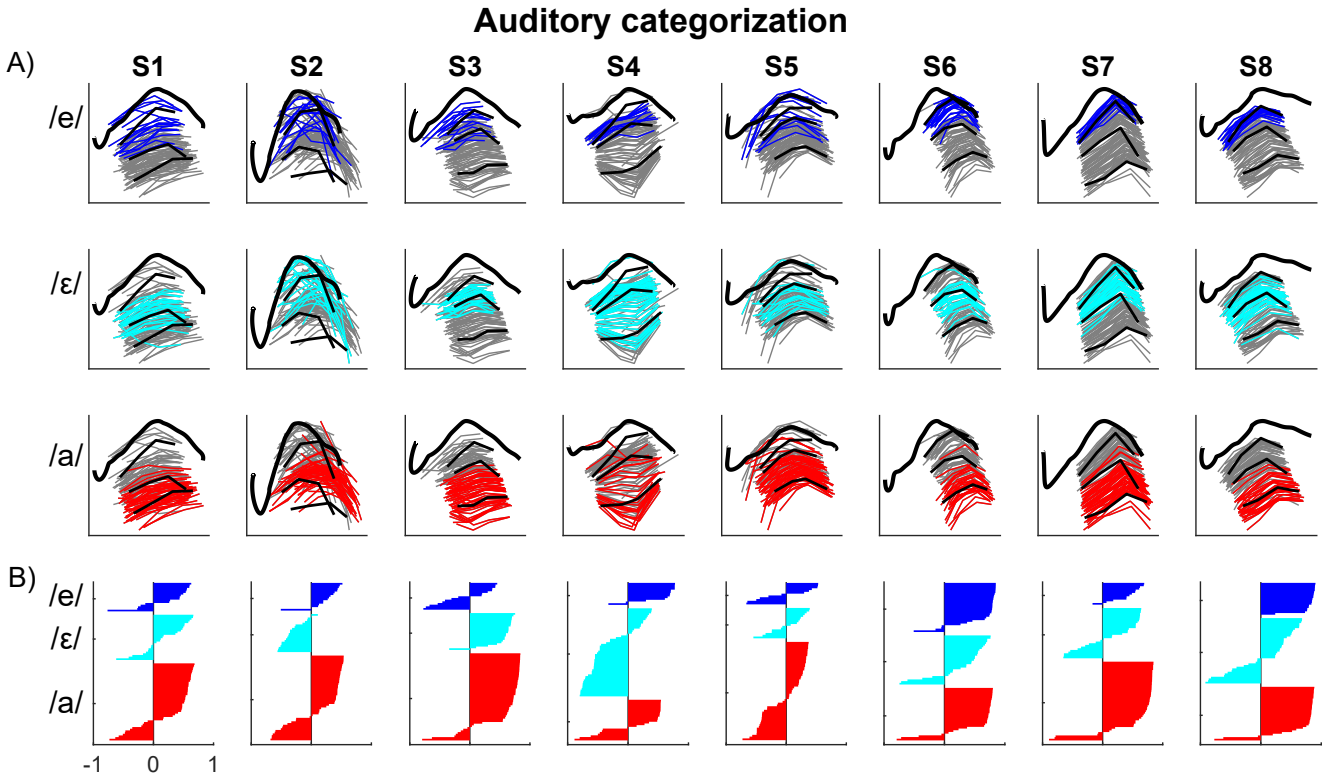

Fig. S5. Reached tongue postures associated to each vowel category according to **auditory labels** (panel A) and associated Silhouette values (panel B) for each subject.

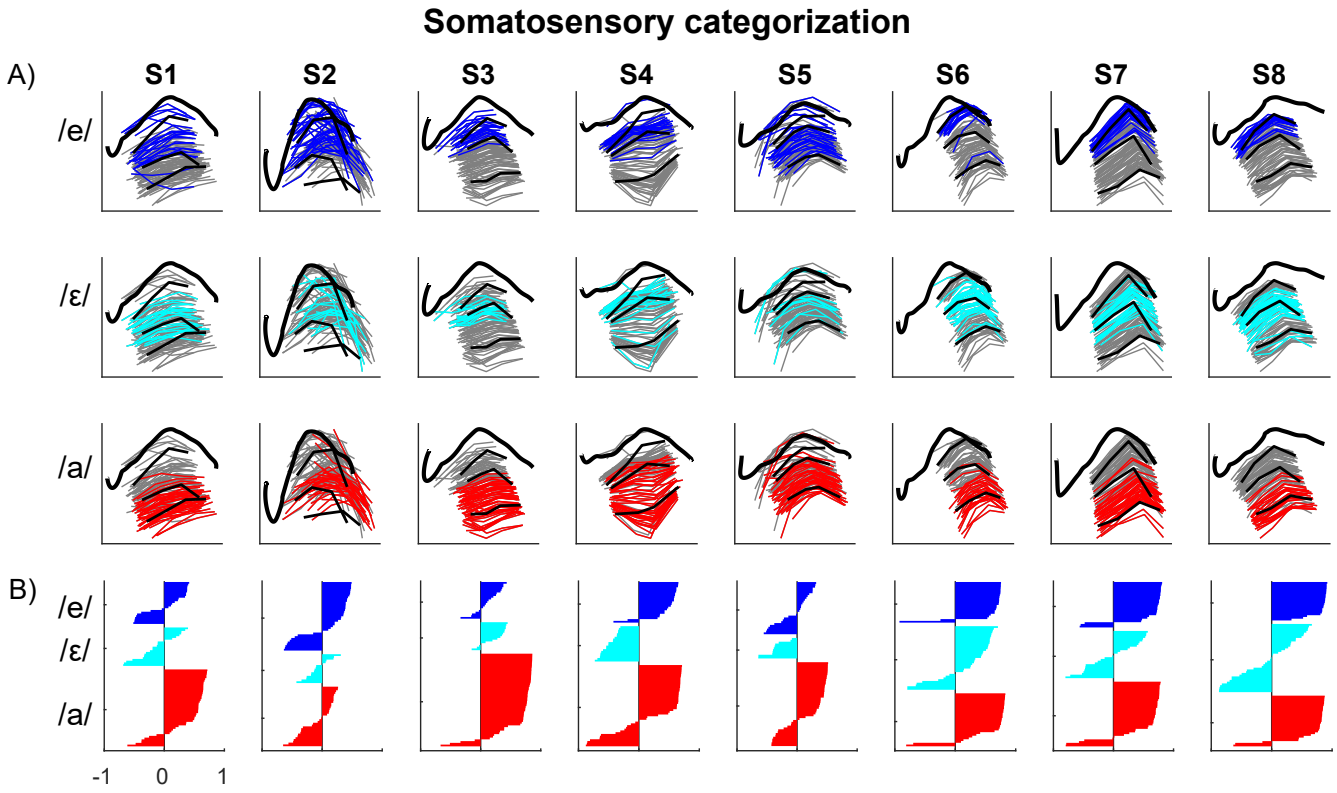

Fig. S6. Reached tongue postures associated to each vowel category according to **somatosensory labels** (panel A) and associated Silhouette values (panel B) for each subject.

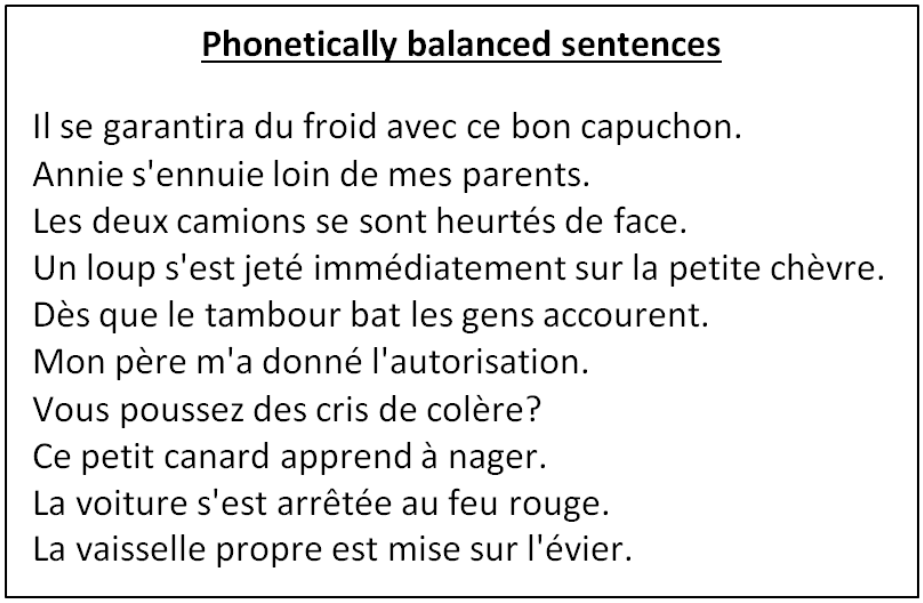

Fig. S7. List of the phonetically balanced sentences used in the habituation task.

References

1. PJ Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53–65 (1987).
2. J Robert-Ribes, JL Schwartz, T Lallouache, P Escudier, Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of french oral vowels in noise. *The Journal of the Acoustical Society of America* **103**, 3677–3689 (1998).
3. P Howell, DJ Powell, Hearing your voice through bone and air: Implications for explanations of stuttering behavior from studies of normal speakers. *Journal of Fluency Disorders* **9**, 247–263 (1984).