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, Pascal Perrier

Speakers are able to categorize vowels based on tongue somatosensation

Jean-François Patri[a,b], David J. Ostry[c,d], Julien Diard[e], Jean-Luc Schwartz[a], Pamela Trudeau-Fisette[f], Christophe Savariaux[a], and Pascal Perrier[a,1]

[a]GIPSA-lab UMR 5216, Grenoble INP, Université Grenoble Alpes, CNRS, F-38000 Grenoble, France; [b]Cognition, Motion and Neuroscience Unit, Fondazione Istituto Italiano di Tecnologia, 16152 Genova, Italy; [c]Department of Psychology, McGill University, Montréal, QC, Canada H3A 1G1; [d]Haskins Laboratories, New Haven, CT 06511; [e]Laboratoire de Psychologie et de Neurocognition (LPNC) UMR 5105, Université Grenoble Alpes, CNRS, F-38000 Grenoble, France; and [f]Laboratoire de Phonétique, Center For Research on Brain, Language, and Music, Université du Québec à Montréal, Montreal, QC, Canada H2X 1L7

Edited by John F. Houde, University of California, San Francisco, CA, and accepted by Editorial Board Member Renée Baillargeon January 27, 2020 (received for review June 28, 2019)

Auditory speech perception enables listeners to access phonological categories from speech sounds. During speech production and speech motor learning, speakers' experience matched auditory and somatosensory input. Accordingly, access to phonetic units might also be provided by somatosensory information. The present study assessed whether humans can identify vowels using somatosensory feedback, without auditory feedback. A tongue-positioning task was used in which participants were required to achieve different tongue postures within the /e, ε, a/ articulatory range, in a procedure that was totally nonspeech like, involving distorted visual feedback of tongue shape. Tongue postures were measured using electromagnetic articulography. At the end of each tongue-positioning trial, subjects were required to whisper the corresponding vocal tract configuration with masked auditory feedback and to identify the vowel associated with the reached tongue posture. Masked auditory feedback ensured that vowel categorization was based on somatosensory feedback rather than auditory feedback. A separate group of subjects was required to auditorily classify the whispered sounds. In addition, we modeled the link between vowel categories and tongue postures in normal speech production with a Bayesian classifier based on the tongue postures recorded from the same speakers for several repetitions of the /e, ε, a/ vowels during a separate speech production task. Overall, our results indicate that vowel categorization is possible with somatosensory feedback alone, with an accuracy that is similar to the accuracy of the auditory perception of whispered sounds, and in congruence with normal speech articulation, as accounted for by the Bayesian classifier.

speech production | somatosensory feedback | categorical perception

Producing speech requires precise control of vocal tract articulators in order to perform the specific movements that give rise to speech sounds. The sensory correlates of speech production are therefore both auditory (associated with the spectrotemporal characteristics of sounds) and somatosensory (related to the position or shape of the vocal tract articulators and to contacts between articulators and vocal tract boundaries). While the propagation of sounds is the means through which linguistic information passes between speakers and listeners, most recent models of speech motor control [DIVA (1), FACTS (2), HSFC (3), Bayesian GEPPETO (4), and ACT (5)] posit that both auditory and somatosensory information is used during speech production for the planning, monitoring, and correction of movements. The crucial role of auditory information has been documented in experiments using bite blocks or lip tubes (6, 7), in which articulation has been shown to be reorganized in order to preserve the acoustical characteristics of speech. The importance of somatosensory information in speech production has been shown in studies in which external perturbations of jaw movement induced compensatory reactions (8, 9). A study of speech production in cochlear implanted patients, who switched their implants on and off (10), has provided evidence that a combination of both sensory inputs results in greater accuracy in speech production.

Auditory speech perception involves the categorization of speech sounds into discrete phonetic units (11, 12), and neural correlates of phonetic representations have been found in the superior temporal gyrus (13). However, it is unknown whether somatosensory information in speech can also be categorized by listeners in a similar and coherent way. In nonspeech tasks, there is extensive evidence of the use of somatosensory information for categorization as, for example, in tactile object recognition (14), but in speech, no study so far has addressed whether speakers are able to identify phonemes based on somatosensory information.

In the present study, we provide evidence that phonological representations can be accessed from somatosensory information alone. Our results indicate that participants are able to recognize vowels without any contribution of auditory feedback, based on tongue postures. This finding has required the design of a paradigm adapted to the specificities of the tongue. Indeed, unlike studies of limb movement in which tests of somatosensory processing can be conducted by using a robotic device to passively displace the limb, provoking passive displacement of

Significance

Auditory speech perception is known to categorize continuous acoustic inputs into discrete phonetic units, and this property is generally considered to be crucial for the acquisition and the preservation of phonological classes in a native language. Somatosensory input associated with speech production has also been shown to influence speech perception when the acoustic signal is unclear, inducing shifts in category boundaries. Using a specifically designed experimental paradigm we were able to demonstrate that in the absence of auditory feedback, somatosensory feedback on its own enables the identification of phonological categories. This finding indicates that both the auditory and orosensory correlates of speech articulation could jointly contribute to the emergence, the maintenance, and the evolution of phonological categories.

Author contributions: J.-F.P., D.J.O., J.D., J.-L.S., and P.P. designed research; J.-F.P., D.J.O., J.D., J.-L.S., P.T.-F., C.S., and P.P. performed research; J.-F.P., D.J.O., J.D., J.-L.S., and P.P. contributed new reagents/analytic tools; J.-F.P., D.J.O., J.D., J.-L.S., and P.P. analyzed data; and J.-F.P., D.J.O., J.D., J.-L.S., P.T.-F., C.S., and P.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. J.F.H. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: Pascal.Perrier@grenoble-inp.fr.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1911142117/-/DCSupplemental.

First published March 2, 2020.

the tongue is very challenging: the tongue is difficult to access inside the oral cavity and highly resistant to displacement. In our paradigm, speakers were instead required to position their tongue using visual feedback in a task that provided no information about the shape and the location of the tongue. We were able to show that subjects succeed in this positioning task, although some speakers are more successful than others. We then tested whether, once they had positioned their tongue, speakers were able to provide a somatosensory categorization of the vowel associated with the reached tongue position. At each reached tongue posture, subjects were asked to whisper in order to enable a subsequent independent auditory evaluation of their articulation by a group of external listeners. The auditory feedback of the speakers was masked by noise in order to ensure that categorization was based on somatosensation only.

We found that speakers were able to identify vowels based on tongue somatosensory information, and there was a good concordance with listeners' judgements of the corresponding sounds. Finally, it is shown that subjects' somatosensory classification of vowels was close to the classification provided by a Bayesian classifier constructed separately from subjects' vowel articulations recorded under normal speech conditions, i.e., with phonation and auditory feedback. These results suggest that phonological categories are specified not only in auditory terms but also in somatosensory ones. The results support the idea that in the sensory–motor representation of phonetic units, somatosensory feedback plays a role similar to that of the auditory feedback.

Results

In order to assess whether vocal tract somatosensory information can be used for phonetic categorization, we first needed a means to instruct subjects to achieve a set of tongue postures without relying on normal speech movement or auditory feedback that might provide nonsomatic cues. We designed a tongue-positioning task using electromagnetic articulography (EMA) to visually guide eight subjects (henceforth somatosensory subjects) toward different target tongue postures, evenly spanning the articulatory range of the three vowels /e, ɛ, and a/ (Fig. 1*B*). Nine target tongue postures were specified for each subject, including three vowel tongue postures corresponding to normal productions of vowels /e, ɛ, and a/, and six intermediate tongue postures distributed regularly over the same tongue vowel workspace. Vowel tongue postures were recorded with EMA for each subject during a preliminary speech production task involving repetitions of vowels under normal speech conditions (*Materials and Methods*). Fig. 1 shows the placement of the EMA sensors and illustrates the set of nine target tongue postures for one representative subject (the set of targets for other subjects are presented in *SI Appendix*, Fig. S3).

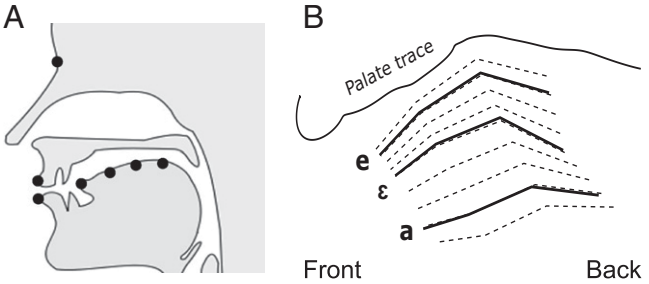

Fig. 1. (*A*) Sagittal view of EMA sensors (black dots) and (*B*) example of target tongue postures from one subject. Targets include three vowel tongue postures (black lines) and six additional intermediate tongue postures (gray dashed lines) distributed regularly over the /e-ɛ-a/ tongue workspace. The solid line on the top of *B* represents the subject's palate trace.

On each trial of the tongue-positioning task (Fig. 2*A*), one of the nine target tongue postures was displayed on a screen under a spatial transformation intended to remove visual information about tongue position and shape (Fig. 2*B*). Target positions were always displayed in the same horizontally aligned configuration at the center of the screen, ensuring that targets in all trials looked the same (red circles in Fig. 2*A*, Bottom). Subjects were provided with real-time visual feedback of their tongue movements according to the spatial transformation shown in Fig. 2*B* and were instructed to 1) move their tongue in order to reach the displayed target within a 5-s time interval (reaching phase), 2) hold the reached tongue posture and whisper with auditory feedback masked by noise (whispering task), and 3) identify the vowel associated with the reached tongue posture (somatosensory identification task).

Importantly, the target tongue postures used in the tongue-positioning task were chosen in order to sample the articulatory workspace from /e/ to /a/ via /ɛ/ as comprehensively and as evenly as possible, in order to obtain a set of well-distributed articulatory configurations for the primary aim of the study, which is to evaluate the use of somatosensory information in the identification of these vowels. Thus, for the tongue-positioning task to be carried out correctly, it was not required that the set of tongue postures which subjects produced at the end of the reaching phase exactly matched the displayed target tongue postures but rather that overall, the set of tongue configurations uniformly sampled the articulatory workspace for the vowels that were studied here.

The sounds that were whispered by the eight somatosensory subjects were identified by eight additional listeners (henceforth auditory subjects) in a forced-choice (/e/, /ɛ/, or /a/) auditory identification task. This perceptual evaluation is crucial since it is the only way to assess whether or not the tongue position described by the EMA sensors corresponded to a vowel and whether or not it was associated with clear auditory characteristics in the /e/–/ɛ/–/a/ vowel range. However, it is also crucial that the whispering phase did not influence the somatosensory identification performed by somatosensory subjects, by providing either auditory or motor cues that might have helped them in their identification task. In regard to possible auditory cues, we have taken a number of precautions in order to minimize the likelihood that subjects could hear themselves during the whispering phase with the auditory feedback masked by noise. To check the effectiveness of this approach, we asked subjects to report whether or not they could hear themselves whisper, and no subject responded that he/she could. We have also evaluated the acoustic power of the whispers and found them to be more than 40 dB below the masking noise, which has been previously shown to make impossible the auditory perception of vowels in French (15) (see *SI Appendix* for details). In regard to possible motor control cues, as is commonly observed in postural control, subjects did not strictly maintain a stable tongue posture during the whispering phase (see *SI Appendix*, Fig. S1, for a quantification of this articulatory variability). To check the possibility that these small tongue movements could have provided helpful motor cues for the somatosensory identification task, we carried out a number of analyses of these articulatory variations, which are described in *SI Appendix*. We found no indication supporting such a possibility. In particular, we found no evidence suggesting that these small movements would be directed toward vowel tongue postures. These observations indicate that the somatosensory identification task was not biased by auditory or motor cues during whispering.

The analysis of the data was divided into two main parts. The first part was devoted to the evaluation of the tongue-positioning task. This first part is necessary for the second part of the analysis since one cannot expect subjects that did not perform the

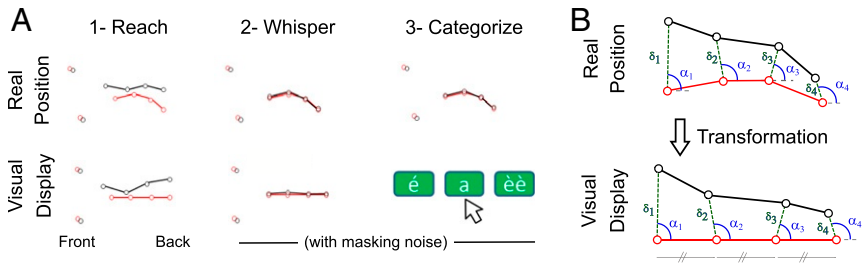

Fig. 2. (*A*) Design of each trial of the tongue-positioning, whispering, and somatosensory identification tasks. (*Top*) The real positions of sensors (sagittal view) corresponding to the target (red circles and lines) and sensors corresponding to subject's tongue (black circles and lines). (*Bottom*) The modified position of sensors as displayed to the subjects. The lip target sensors were displayed as vertically aligned circles in the left part of the display and were intended to help subjects to keep their lip position constant during the task. (*Bottom Right*) The three alternative forced choice display of the somatosensory identification task. (*B*) Illustration of the spatial transformation used for the visual display in the tongue-positioning and whispering task. Red circles and lines correspond to the target tongue posture and black circles, and lines correspond to the actual tongue shape. The target tongue posture is transformed in such a way that all segments become equal in length and horizontally aligned. Then the actual tongue shape is deformed by preserving δ_i and α_i after transformation.

tongue-positioning task correctly to succeed in the somatosensory identification task. More specifically, we assessed whether the participants reached tongue postures 1) that varied continuously and smoothly over the whole /e, ε, and a/ articulatory range and 2) that corresponded to sounds that can be reliably auditorily identified as a vowel in the /e, ε, and a/ range. Tongue postures meeting these evaluation criteria will henceforth be referred to as vowel-like tongue postures.

In the second part of the study we assessed performance in the somatosensory identification task. This had three main parts. First we evaluated the separability of the three vowel categories as obtained by somatosensory categorization. Second, we compared the somatosensory categorization with auditory categorization provided by the auditory subjects who evaluated the whispered sounds. Finally, we compared the somatosensory categorization with the outcome of a Bayesian classifier that relied on tongue postures recorded during normal speech movements.

Stage 1: Evaluation of the Tongue-Positioning Task.

Articulatory analysis of the tongue-positioning task. Fig. 3 shows for each of the participants in the study the set of tongue postures reached in the tongue-positioning task, superimposed on the average tongue configurations measured for /e, ε, and a/ during normal speech production. It can be seen that for subjects S3, S6, S7, and S8 the set of reached tongue postures (Fig. 3, *Bottom*, left-hand side of each panel) uniformly covers the /e, ε, and a/ range, whereas for the remaining subjects (S1, S2, S4, and S5; Fig. 3, *Top*, left-hand side of each panel), there were noticeable deviations from the average vowel postures. In order to quantitatively assess this observation, we first evaluated whether the set of reached tongue postures actually covered the expected /e, ε, and a/ range of tongue configurations. We also assessed quantitatively whether these tongue postures were uniformly distributed over the range and direction associated with the set of target tongue postures. We conducted two separate principal component analyses (PCAs) for each subject separately, one using their set of target tongue postures and the other using their set of reached tongue postures. Details about this analysis are provided in *SI Appendix*. We summarize the main results below.

The PCA on the nine target tongue postures showed that the target tongue posture workspace was well described by a single dimension for all subjects but S2. The second PCA showed that only subjects S3, S6, S7, and S8 produced reached tongue postures that were also well represented by a single dimension for the reached tongue postures (see the two clusters in Fig. 3*B*). Moreover, for these four subjects we also found a good match between the single direction which characterized the target tongue postures and the one which characterized the reached tongue postures. We also 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. In each panel of Fig. 3*A* the right side shows the resulting distribution over the range of the nine target tongue postures. For all subjects apart from S1 and S5 we observe quite

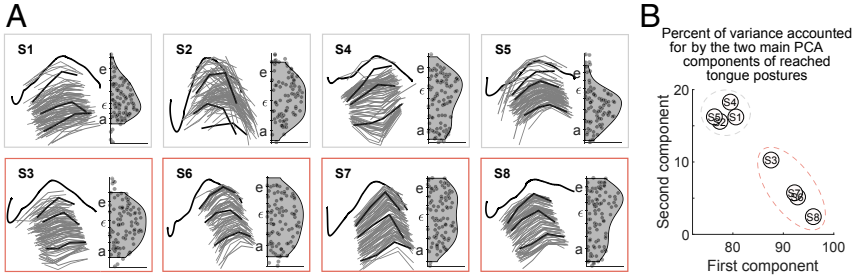

Fig. 3. (*A*) Distribution of the set of 90 tongue postures reached by each somatosensory subject across trials in the tongue-positioning task. For each subject the set of reached tongue postures (gray lines) is represented on the left-hand side of each panel, together with the average /e, ε, a/ tongue postures (black lines) obtained from the speech production task (*Materials and Methods*). The right-hand side of each panel presents the distribution of reached tongue postures in the main /e, ε, a/ direction (represented vertically), as determined by a PCA carried out on the set of target tongue postures. (*B*) Clustering of the eight subjects in two classes (elliptical contours drawn by hand), based on the proportion of variance explained by the first two principal components describing the set of reached tongue postures.

uniform distributions, and this was confirmed by a statistical Kolmogorov–Smirnov test.

Auditory analysis of the adequacy of the reached tongue postures. In order to investigate the adequacy of the reached tongue postures for purposes of somatosensory vowel identification, we evaluated whether the whispered sounds associated with these postures were compatible with vowel production. To do so, subjects' whispered productions were classified by a separate group of listeners (auditory subjects). First, we analyzed the consistency of the responses provided by auditory subjects in the auditory identification task, which we call auditory answers henceforth. Second, we inferred from the set of auditory answers given for each whispered sound a canonical auditory label for this sound (see *Materials and Methods* for details) and assessed whether these auditory labels were associated with well-separated clusters of tongue postures.

Consistency of Auditory Answers Across Auditory Subjects. We assessed the consistency of the auditory classification of each reached tongue posture recorded during the tongue-positioning task by computing the entropy of the distribution of auditory answers attributed to each of the whispered utterances (see *SI Appendix* for details). We expected that whispers which sound like whispered vowels would be labeled consistently by auditory subjects and would therefore result in distributions of auditory answers with low entropy. On the other hand, we expected that whispers that would not sound like whispered vowels would be labeled with greater uncertainty, resulting in distributions of auditory answers with greater entropy (close to uniformly random answers). Fig. 4 presents violin plots of the distribution of entropies of auditory answers for each whispered sound produced by each of the somatosensory subjects. From this figure it can be seen that the whispers of all but two somatosensory subjects (S4 and S5) have low average entropy, with violin plots being larger at the base than at the top. This means that most of the whispers produced by the somatosensory subjects were classified in a consistent way across auditory subjects. Statistical analysis confirmed that the distribution of entropy differs across subjects (Kruskal–Wallis $\chi^2 = 123.95$, $P < 0.001$). Pairwise comparisons revealed that the entropy distributions for subjects S4 and S5 were significantly greater than for the others ($P < 0.01$ with Wilcoxon test and Benjamini and Hochberg correction method for multiple comparisons). This indicates that the whispers produced by these two subjects conveyed little information about vowel identity, presumably because the associated reached tongue postures did not correspond to one of the sounds /e, ɛ, and a/ that were under test.

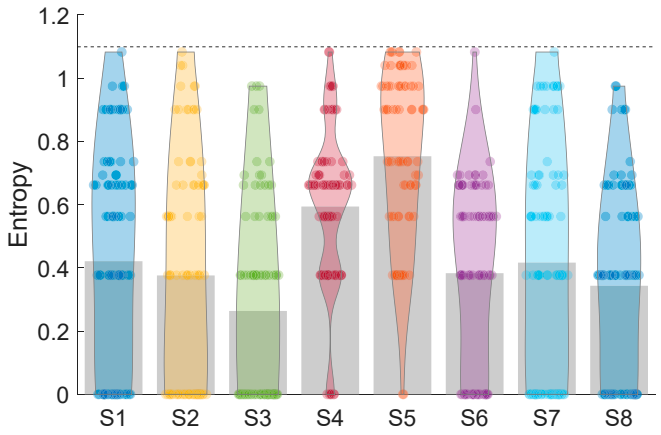

Fig. 4. Distribution of entropy of auditory answers for each whisper produced by each somatosensory subject (violin plots). Small entropy values correspond to whispers with low uncertainty about vowel identity; in particular, zero entropy values correspond to whispers that were given the same label by all auditory subjects. Data points correspond to the different whispers performed across trials by the somatosensory subjects during the positioning task. They are distributed over 10 entropy values, which correspond to possible entropy values for the three vowel labels and eight auditory answers (see *SI Appendix* for details). The dashed line indicates the theoretical maximum uncertainty (entropy of 1.1) of auditory labeling. Gray colored bars represent the average entropy of the auditory answers for each somatosensory subject.

Consistency and Separability of the Clusters of Reached Tongue Postures Associated with the Three Auditory Labels. In normal speech production there is substantial similarity of tongue posture across repetitions of the same vowel (consistency across repetitions), and these tongue postures can also be reliably distinguished from the tongue postures associated with repetitions of another vowel (separability across vowels). In order to check whether the tongue-positioning task reproduced these speech characteristics, we asked whether or not the set of tongue postures associated, for example, with sounds that listeners judged as the vowel /a/ was different from the set of tongue postures associated with sounds that listeners judged as /ɛ/. We assessed the auditory labels assigned by the set of auditory subjects by evaluating the consistency and the separability of the grouping of tongue postures made on the basis of the auditory labels. We expected that whispers that carry relevant information for vowel classification should be associated with articulatory characteristics that are 1) consistent within each category and 2) different enough across categories to preserve their distinctiveness. Hence, if the four EMA sensors are good descriptors of the posture of the tongue, the auditory labels for the whispers that sound like one of the /e, ɛ, and a/ vowels should be associated with quite compact and well-separated clusters of reached tongue postures. In the case of whispers that do not sound like one of these vowels, the clusters of reached tongue postures should be wide (more variable) and largely overlap one another. Silhouette scores provide a measure of clustering quality in terms of consistency and separability, by comparing how well each data point belongs to its own cluster as compared to its neighboring clusters (16). Our silhouette analysis assigns to each tongue posture a value in the $[-1, 1]$ range, with positive values corresponding to data that are well inside their cluster, close to 0 values corresponding to data that are close to the boundaries of their cluster and negative values corresponding to data that are closer to a neighboring cluster than to their own cluster (see *SI Appendix* for details). The gray bars in Fig. 5, *Middle*, show for each somatosensory subject the average silhouette score of clusters of reached tongue postures based on auditory labels. We call these scores auditory clustering scores. They are arranged in Fig. 5 in descending order from left to right. Fig. 5, *Middle*, also shows average silhouette scores of clusters of reached tongue postures based on somatosensory labels (see below for an analysis of somatosensory clustering). Examples of tongue clusters associated with high and low average silhouette scores are shown in the left and right upper panels in Fig. 5, respectively (figures corresponding to other subjects are presented in *SI Appendix*, Fig. S5). It can be seen that for a subject with high silhouette scores (S6; Fig. 5, *Left*), the clusters of tongue postures are well separated with moderate overlap between auditory vowel categories. In contrast, clusters of a subject with low silhouette scores (S5; Fig. 5, *Right*) show strong overlap. Furthermore, for well-separated clusters, the tongue postures associated with the auditory labels /e/, /ɛ/, and /a/ are correctly distributed from high to low around the tongue postures characteristic of each vowel (black lines in Fig. 5, *Left*).

In order to assess the significance of the auditory clustering scores, as compared to the null hypothesis that tongue postures were labeled at random, we performed a nonparametric

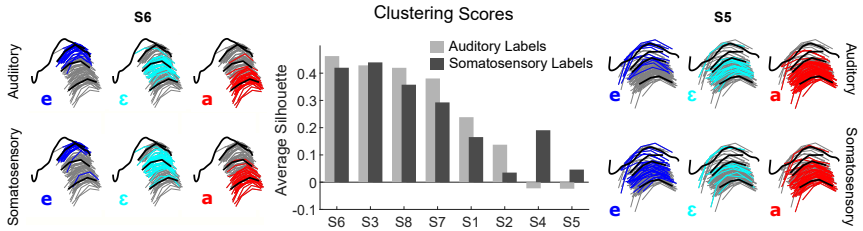

Fig. 5. Clustering of the reached tongue postures associated with auditory and somatosensory labels. (*Middle*) Bars show the auditory and somatosensory clustering scores obtained for each somatosensory subject, arranged in descending order (from left to right) of auditory clustering scores (light gray bars). Auditory clustering scores are significantly different from chance ($P < 0.01$) for all subjects except S4 and S5. Somatosensory clustering scores are significantly different from chance ($P < 0.01$) for all subjects. The clusters of reached tongue postures associated to each vowel category for a representative subject corresponding to (*Left*) good and (*Right*) poor clustering scores ([*Top*] auditory scores and [*Bottom*] somatosensory scores). For each vowel category, the associated target tongue postures are specifically colored in order to distinguish them from the other reached tongue postures displayed in gray (as in Fig. 3). Black lines correspond to the three vowel tongue postures indicated for reference (upper, middle, and bottom black lines corresponding to /e/, /ɛ/, and /a/, respectively).

randomization test (17) with 10^4 random permutations. It was found that auditory clustering scores were significantly different from chance ($P < 0.01$ with Holm–Bonferroni correction for multiple comparisons) for all subjects except S4 and S5. It can be seen that these two subjects also have the highest values of entropy of the auditory answers provided by listeners (Fig. 4). This is consistent with our hypothesis, stated above, that their whispers conveyed little relevant information for vowel identification because their reached tongue postures poorly correspond to configurations compatible with vowel production. Two other somatosensory subjects (S1 and S2) had silhouette scores that were clearly lower than those of the remaining four subjects, despite the fact that their whispers seemed to convey some relevant information for vowel identification, as indicated by their low level of entropy of the auditory answers (Fig. 4). In contrast, the high auditory clustering scores of the remaining four somatosensory subjects (S3, S6, S7, and S8), for whom low entropy auditory classification was also observed, indicate that these subjects regularly reached tongue postures compatible with typical productions of the vowels /e/, /ɛ/, and /a/ and that the regularities observed in the EMA sensor positions are reliable indicators of the whole tongue posture typically associated with these vowels. This interpretation of the relationship between tongue postures and sounds is further supported by the fact that the three subjects with the highest clustering scores are also those who present the fewest data points with high entropy in the violin plots of Fig. 4.

Stage 2: Evaluation of Performance in the Somatosensory Identification Task. We next focus on subjects' classification based on their somatosensory feedback. The analysis addresses three specific questions: 1) Was each somatosensory subject able to use somatosensory feedback to identify which vowel among /e/, /ɛ/, and /a/ was associated with each reached tongue posture? 2) Was somatosensory vowel identification consistent with auditory identification, providing specific categorization regions well separated from one vowel category to the other? 3) Was somatosensory vowel identification consistent with articulatory patterns?

The previous analysis showed that there are important differences across subjects in their ability to perform the tongue-positioning task, both in terms of coverage of the /e, ɛ, and a/ task workspace and in the production of conventional tongue postures associated with these sounds. In the evaluation of the somatosensory identification, which follows, we will focus on the subjects who performed well in the tongue-positioning task because they satisfactorily reached vowel-like tongue postures, as evidenced by stage 1 of our analysis (Figs. 3–5): subjects S3, S6, S7, and S8. The results obtained for the other subjects (S1, S2, S4, and S5) will be also displayed for sake of information only.

We firstly evaluated somatosensory identification by analyzing the consistency and the separability of the clusters of tongue postures associated to each vowel category on the basis of the answers provided in the somatosensory identification task (we call these answers somatosensory labels henceforth). As in auditory identification, relevant somatosensory classification is expected to be associated with quite compact and well-separated clusters, and irrelevant classification is expected to be associated with wider and poorly separated clusters. The black bars in Fig. 5, *Middle*, present the average silhouette values associated with the somatosensory labels of each somatosensory subject (see *SI Appendix*, Fig. S6, for the clustering of tongue shapes for each subject and associated silhouette values). We call these scores somatosensory clustering scores. We assessed whether or not the somatosensory labeling was random using nonparametric randomization tests similar to those used for auditory clustering scores (10^4 random permutations). We found that the labeling was different from chance for all subjects ($P < 0.01$ with Holm–Bonferroni correction for multiple comparisons). Subjects S3, S6, S7, and S8, who performed well in the tongue-positioning task and had the highest auditory clustering scores, likewise had auditory and somatosensory clustering scores that were similar in magnitude. This indicates that these four subjects were able to categorize their own tongue postures on the basis of somatosensory information as efficiently as the auditory subjects who classified their whispered speech. As expected, subjects who did not perform the tongue-positioning task well also had lower somatosensory clustering scores.

For each somatosensory subject we constructed a somatosensory and an auditory identification curve along an articulatory continuum that spanned the nine target tongue postures. To do so, we subdivided the set of reached tongue postures into nine subsets corresponding to each of the nine target tongue postures: each subset was composed of the set of reached tongue postures that were closest to the associated target in articulatory space. This resulted in nine steps along an articulatory continuum from /e/ to /a/. In each subset of reached tongue postures associated with the steps along this continuum we computed the proportion of tongue postures that were classified by speakers or listeners as each of the vowels /e/, /ɛ/, and /a/ (see *SI Appendix* for details). We quantified the similarity of somatosensory and auditory identification curves by computing Pearson's correlation coefficient between them. Fig. 6 illustrates the resulting identification curves and the corresponding correlation coefficients for each of the somatosensory subjects. Fig. 6, *Left*, presents the curves for the subjects who performed well the

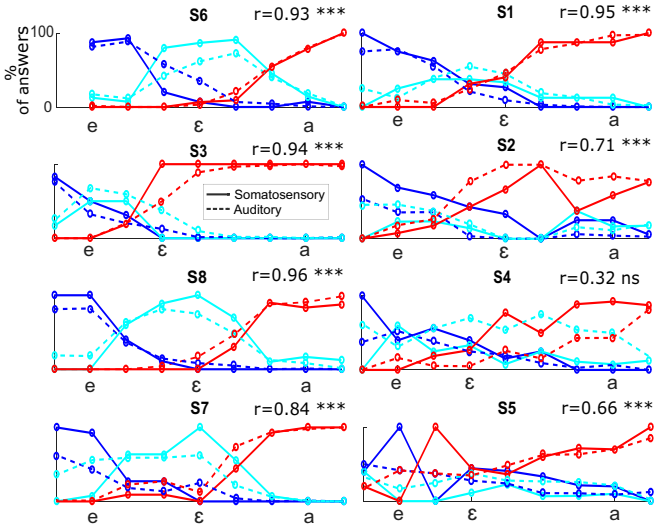

Fig. 6. Identification curves of somatosensory labels (plain lines) and auditory labels (dashed lines) for subjects associated with (*Left*) good and (*Right*) poor clustering scores. Subjects are arranged within *Left* and *Right* in descending order (from the top to the bottom) of auditory clustering scores, as in Fig. 5. Blue, cyan, and red curves correspond to /e/, /ɛ/, and /a/ categories, respectively. For each of these curves the nine dots along the *x* axis indicate the percentage of answers for the corresponding vowel for each of the nine subsets of reached tongue postures. In all of the panels, Pearson's correlation coefficient is reported as a measure of similarity between the auditory and somatosensory identification curves. Statistical significance: ns = nonsignificant; *** = significant ($P < 0.001$).

tongue-positioning task, and Fig. 6, *Right*, gives the curves for the other subjects.

Fig. 6, *Left*, shows curves that display a clear separability of the regions associated with each vowel in the somatosensory space. First, in terms of similarity, the somatosensory identification curves are very close to the auditory identification curves (Pearson's correlation coefficient above 0.8, $P < 0.001$ from two-tailed Student statistics implemented in the MATLAB functions corr, with Holm–Bonferroni correction for multiple comparisons) that, in agreement with conventional observations, display well-separated acoustic regions for each vowel. Second, we clearly see in each curve the existence of three distinct categories with regions in which only one vowel gets high identification scores and in a sequential order that matches the sequential order along the articulatory continuum. As expected, this pattern is not observed in the curves obtained from the subjects that did not perform the tongue-positioning task well (Fig. 6, *Right*).

We further explored the consistency between vowel identification provided by subjects in the absence of auditory feedback and the associated tongue postures by comparing the somatosensory identification curves with identification curves resulting from a Bayesian classifier that models the link between vowel categories and tongue postures during normal speech production (see *SI Appendix* for details). We determined the parameters for this classifier from the statistical distribution of the tongue postures reached for each vowel during the preliminary production task. We used the model to predict the identification curves that would be expected for each subject if their classification was based on statistics of the tongue postures that they achieved during normal speech production (see *SI Appendix* for details).

Fig. 7 presents the identification curves predicted by the Bayesian classifier (dashed lines) along with those obtained from the actual somatosensory identification (solid lines). For the four subjects who performed well in the tongue-positioning task (Fig. 7, *Left*), there is a clear similarity between the two curves as reflected by the high Pearson correlation coefficients. The close agreement between these two curves provides further evidence of a strong link between tongue postures achieved during normal speech production and the perception of these vowels based on somatosensory feedback. As expected, for the remaining four subjects shown Fig. 7, *Right*, the similarity between curves is less.

Discussion

Somatosensory Information May Provide a Perceptual Basis for Vowel Categories. We have designed an experimental task to test whether speakers have access to phonologically relevant somatosensory information from the tongue. In this task subjects were required 1) to reach a range of vowel-like tongue postures between the vowels /e-ɛ-a/, in a setup designed to minimize the use of everyday speech motor control strategies, and 2) to identify the associated vowels on the basis of tongue somatosensory feedback.

A first step in the data analysis involved an evaluation of whether subjects were able to correctly perform the tongue-positioning task. In order to do so, we assessed 1) whether subjects were able to produce a range of articulations that covered the intended vowel space and 2) whether subjects reached tongue postures that correspond to the production of a vowel, by evaluating the consistency of listeners' perceptual judgments. Despite the difficulty of the task, four subjects (S3, S6, S7, and S8) were clearly able to perform the task correctly: 1) they achieved tongue postures that were well distributed over the entire intended vowel space (Fig. 3); 2) their whispers were well identified as vowels by independent listeners, and these identifications were consistent across listeners (Fig. 4); and 3) the clusters of reached tongue postures associated with each auditory vowel category were only partially overlapping and consistent with the expected vowel articulations (Fig. 5, gray bars). The remaining four speakers (S1, S2, S4, and S5) either failed to achieve correct tongue postures, were too variable, or produced whispered speech that was not correctly classified by listeners in the auditory identification task. We interpret these observations as evidence for the fact that for these latter four subjects, tongue postures as a whole did not correspond to vocal tract configurations that are appropriate for vowel production. As a consequence we focused on subjects S3, S6, S7, and S8 for the evaluation of their capacity to identify vowels

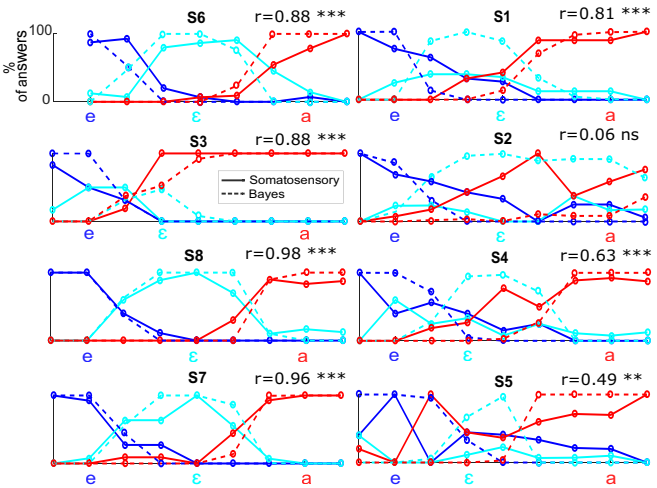

Fig. 7. Identification curves from somatosensory labels (plain lines) along with predictions from a Bayesian classifier (dashed lines). In all panels, Pearson's correlation coefficient is reported as a measure of similarity between model predictions and somatosensory identification curves. Statistical significance: ns = nonsignificant; ** = significant ($P < 0.01$); *** = significant ($P < 0.001$). See Fig. 6 for more details.

based on somatosensory feedback. We found that these four subjects were quite successful in the somatosensory identification task since they judged vowel categories in a consistent way and they did so in a way that matched the identification of their whispered sounds by auditory subjects (Fig. 5). Hence, a first important result of our study is that 100% of the participants who, according to the criteria that we described above, successfully performed the tongue-positioning task also succeeded in the somatosensory identification of the three vowels of interest.

These subjects, who produced consistent vowel classification based on their own tongue postures, yielded somatosensory identification curves across the /e-ɛ-a/ continuum that were similar to the ones provided by auditory subjects listening to their whispers. Moreover, a comparison of the somatosensory and auditory identification curves shows that for these subjects, somatosensory identification closely follows the categorical pattern of auditory identification. We also find that their somatosensory identification curves are close to those of a Bayesian classifier that provides a categorization of each of the vowels of interest based on statistics of the tongue postures achieved during normal speech production. Hence, our study provides evidence that somatosensory feedback on its own provides relevant information for vowel identification; that this identification has categorical properties, at least along the /e-ɛ-a/ direction from high-front to low-back; and that there is a strong link between perception based on somatosensory feedback and tongue postures produced in natural speech.

Note that the conclusion that speech phonetic categories can be perceptually identified from somatosensory information is not weakened by the inability of some subjects to successfully perform the tongue-positioning task. Successful performance of the task was a prerequisite for us to evaluate somatosensory identification in our participants. The tongue-positioning task required that subjects perform a complex visuo-motor-somatosensory match, which was needed to avoid providing movement cues that might be used for identification and because passive positioning the tongue is effectively impossible. Not surprisingly only half of our subjects could perform this task satisfactorily.

How Does Somatosensory Feedback Operate in Vowel Perception? The role of somatosensory information in speech perception has been previously demonstrated in several studies. Using a robotic device to stretch the facial skin, Ito et al. (18) showed that the perceptual boundary between vowels /ɛ/ (in "head") and /æ/ (in "had") was altered when the stretch was applied with a timing and in directions compatible with the somatosensory input associated with the production of these speech sounds. These results are in accord with the findings of Fowler and Dekle (19), who have shown that the perceptual boundary between the labial stop /b/ and the alveolar stop /d/ on a /ba-da/ acoustic continuum was influenced by haptic information which resulted from a speaker silently mouthing one of these two syllables in synchrony with the acoustic stimulus. In the same vein, Gick and Derrick (20) applied slight inaudible air puffs to the neck or the hand of participants during a perceptual identification test of /ba/ and /pa/ stimuli in noise. They observed that the presence of the air puffs increased the perception of the unvoiced /pa/, whether it was applied to the neck or to the hand. Thus, information directly or indirectly linked with somatosensation during speech production contributes to speech perception. More generally, these studies support the idea that speech perception integrates all sensory inputs that are associated with speech sounds, both those that are associated with the normal experience of speech and those that are due to specific training (see, for example, the enhancement of lipreading of speech sentences due to the application of vibrotactile vocoders to the lips in ref. 21). However, so far, the contribution of somatosensory feedback to speech perception was only shown to be modulatory. The present study shows that somatosensory information also plays a linguistic role in speech perception; namely, it provides relevant information for phonetic categorization. The existence of a somatosensory route to speech perception could be attributable to different cognitive processes. First, in the theoretical framework according to which there are targets in the sensory domains that specify goals for the production of the phonetic units, the identification of phonetic units in the absence of auditory feedback could directly occur in the somatosensory domain, with regions of this domain typically associated with these units, as was proposed for example by Keating (22) or Guenther et al. (23). This might come about through the repeated pairing of speech sounds and associated somatic feedback during speech production and speech learning. By this account, one would expect that somatosensory classification would remain possible even if auditory cortex was suppressed.

Also in the context of the existence of sensory targets for phonetic units, an alternative explanation would involve a sensory-to-sensory map in the brain, as suggested by Tian and Poeppel's Dual Stream Prediction Model (24, 25) developed in the context of studies of mental imagery in speech production and perception using magnetoencephalography. These authors hypothesized that abstract auditory representations could be predicted on the basis of estimated somatosensory consequences of motor commands. In this context, somatosensory vowel identification would result from the association of somatosensory characteristics with auditory phonological categories, via a sensory-to-sensory map. By this account, vowel categorization would in fact occur in the auditory domain rather than in the somatosensory domain. Accordingly, somatosensory classification should not be possible without the involvement of auditory cortex.

Alternatively, as was proposed by Houde et al. (26) and more recently by Parrel et al. (2), phonetic categorization could arise from the estimation of the state of the vocal tract described in terms of phonologically relevant geometrical parameters. This estimation, which establishes a link to phonetic categories, uses auditory and somatosensory feedback, not for a comparison with sensory targets but to evaluate whether the vocal tract state, predicted on the basis of the motor commands, corresponds to the actual state of the vocal tract and if necessary to update this prediction. This might be tested by perturbing the cerebellum, which is widely assumed to be crucial for feedforward control using forward models (27), with the prediction that somatosensory phonetic categorization should be rendered impossible.

By showing that there is categorical somatosensory decoding of speech configurations, the present study is intrinsically compatible with the various cognitive processes that are postulated above, but it does not enable one to disentangle the alternatives. Nevertheless, the present study demonstrates that somatic information plays a role at a linguistic level in addition to its previously documented involvement in processes of speech production and learning.

Materials and Methods

Participants. The overall study was composed of two parts that involved two different groups of subjects (eight subjects, four males, in each group, ages ranging between 20 and 36 y, average 24 y, for the first group and between 21 and 35 y, average 29 y, for the second group). The first group of subjects, called somatosensory subjects, performed all tasks described below except the auditory identification task. The second group of subjects, called auditory subjects, performed only the auditory identification task. All subjects were native French speakers and reported no cognitive or hearing impairment.

The experimental protocol was conducted in accordance with the ethical standards specified in the 1964 declaration of Helsinki and was approved by the institutional ethics committee of the University Grenoble-Alpes (IRB00010290-2016-07-05-08). All subjects provided informed consent.

Experimental and Tasks Design.

Overview. The experimental protocol was designed to achieve two goals. The first goal was to have subjects produce a set of tongue configurations that covers the articulatory range of the French vowels /e, ɛ, a/. These particular vowels were selected for this study because their primary articulatory features can be readily observed with an EMA system, which provides highly accurate information in the front part of the vocal tract (from the lips to the region of the soft palate). The second goal was to minimize the likelihood that subjects use anything other than somatosensory information, such as auditory or visual information or stored motor strategies, to identify the vowel associated with their tongue configurations.

In a preliminary production task we used EMA to record for each subject tongue postures associated with the production of vowels /e/, /ɛ/, and /a/. Using the average tongue posture for each vowel, we defined a set of intermediate postures by linear interpolation and extrapolation along /e–ɛ/ and /ɛ–a/, such that overall a uniformly distributed set of nine target tongue postures was obtained, with seven of them spanning the articulatory space from /e/ to /a/ average tongue postures and the last two extrapolated outside of this range (one beyond /e/, the other beyond /a/). An example of the set of target tongue postures is shown in Fig. 1, *Right* (for the remaining subjects, see *SI Appendix*, Fig. S3). This set of target tongue postures was then used in the tongue-positioning task described below.

In the tongue-positioning task, subjects (somatosensory subjects) could see a real-time visually distorted display of the EMA sensors on their tongue, connected by solid lines, along with a distorted static display of the intended target. They were given 5 s (reaching phase) to move their tongue toward the displayed target.

The visual display was designed to avoid providing the subject with visual information about tongue position or shape (Fig. 2*A*). Target positions were always displayed in the same horizontally aligned configuration at the center of the screen, ensuring that targets in all trials looked the same (red circles in Fig. 2*A*). Then the position and movement of the subject's sensors (black circles in Fig. 2*A*) were displayed on the screen relative to the position of the displayed target according to the spatial transformation shown in Fig. 2*B*. When the tongue was in the target tongue posture, the sensors' position matched the horizontally displayed target configuration in the visual display. In other words, at the target tongue posture, the displayed position of subject's EMA sensors did not correspond to the actual physical configuration of the subject's tongue but rather was warped onto a straight line. In this way, visual information about actual tongue shape and position was eliminated (see real position vs. visual display panels in Fig. 2*A*). The tongue-positioning task thus minimized visual cues regarding tongue posture and the availability of speech motor control strategies. The goal was to ensure that the subsequent judgment of tongue postures (the primary task) was not based on anything other than somatosensory information.

After the reaching phase of the positioning task, somatosensory subjects were instructed to whisper while holding the reached posture (whispering task). In-ear masking noise (white noise) was used in the whispering task to prevent subjects from identifying their reached tongue postures based on auditory information (Fig. 2*A*, *Middle*). Masking noise was sent through in-ear headphones, compatible with the EMA system (Etymotic ER1). After each positioning and whispering trial, subjects performed a somatosensory identification task by indicating which vowel they thought they were whispering. The response was a three-alternative forced choice among the target vowels /e/, /ɛ/, and /a/, as illustrated in Fig. 2*A*, *Bottom Right*. The masking noise was maintained during the somatosensory identification task.

Subjects' whispers were classified by an additional group of listeners (auditory subjects) in an auditory identification task. Auditory subjects were instructed to indicate which vowel corresponded best to the whispered sound using a three-alternative choice design similar to the one in the somatosensory identification task. Each subject in this auditory identification task heard all whispers produced by all somatosensory subjects, and each whisper was heard and identified only once by each subject.

The task performed by the somatosensory subjects involved two sessions that were completed on two consecutive days. The session on the first day was set up in order to provide subjects with training for the positioning and whispering task. This tongue-positioning and whispering task was preceded by a production task under normal condition, which was used for the definition of the target tongue postures used in tongue positioning and whispering. The session on the second day again included the production task under normal conditions for the definition of new target tongue postures, along with the tongue-positioning and whispering tasks, which used these new target tongue postures, and the somatosensory identification task. These last two tasks are the main experimental tasks of interest. No use was made on the second day of any of the data recorded on the first day. The first day was strictly for training. Importantly, no reference to the forthcoming somatosensory identification task was given during the training on the first day, in order to prevent the possible development of strategies for the tongue-positioning task. In all tasks, subjects were seated in a soundproof room in front of a computer screen. The influence of the jaw on tongue positioning was removed by means of a bite block specifically molded for the teeth of each subject (Coltene PRESIDENT putty). The bite block was designed to keep the jaw in a comfortable position but to ensure that articulatory positioning was restricted to the tongue. The bite block was molded by the subject while holding a reference spacing of 15 mm between their teeth.

Design of specific tasks. The sessions on both days began with a habituation task followed by a production task in which specific tongue postures were recorded and used for each subject separately to establish the target tongue postures that were used in the positioning and whispering tasks which followed. The habituation task was performed right after the placement of the EMA sensors and consisted of the production of 10 phonetically balanced sentences intended to get subjects used to speaking with the sensors glued to the tongue (see *SI Appendix*, Fig. S7, for the list of phonetically balanced sentences). The bite block was not used in this habituation task. After the habituation task, the production task consisted of the normal production of five French vowels (/e/, /ɛ/, /a/, /ɔ/, and /œ/). In each trial of the production task the subjects started from the tongue posture for the consonant /l/ (see below). This was followed by a visually displayed "go" signal, upon which subjects were required to produce the /l/ sound followed by one of the five French vowels (/e/, /ɛ/, /a/, /ɔ/, and /œ/). In order to maintain a certain level of naturalness, the consonant–vowel (CV henceforth) syllables were displayed as French words ("les" [plural "the"] for /le/, "lait" ["milk"] for /lɛ/, "la" [feminine "the"] for /la/, "lors" ["during"] for /lɔ/, and "leur" ["their"] for /lœ/). Subjects were instructed to only produce the syllable onset and nucleus and to hold the vowel for around 3 s (with the help of a countdown displayed on the screen). Each of the five CV items was produced 10 times in random order. The production task and all following tasks were performed with the bite block.

We selected the consonant /l/ for the onset of the syllable because we wanted all of the five vowels to be articulated in a prototypical way, in order to have typical tongue postures for each vowel. This last point required that the selected initial consonant should have little influence on the articulation of the subsequent vowel. The French consonant /l/ has an apical, nonvelarized articulation location. A number of studies (summarized in table 1 of ref. 28) have shown that the nonvelarized /l/ enables a fair amount of variability in the anterior–posterior positioning of the tongue associated with the articulation of the subsequent vowel. In addition, electropalatographic data (e.g., ref. 29, p. 189) provided evidence that the nonvelarized /l/ does not feature tongue grooving but an arched tongue, in the coronal plane. These observations suggest that consonant /l/ should minimally constrain the articulation of the subsequent vowel, contrary, for example, to the postalveolar sibilants /s/ and /ʃ/, which would introduce articulatory constraints in terms of anterior–posterior positioning and/or coronal tongue grooving.

The set of target tongue postures for the tongue-positioning task was obtained from average tongue postures produced for vowels /e/, /ɛ/, and /a/ during the production task. Average tongue postures were obtained from the 10 repetitions of each vowel. Six intermediate target tongue postures were computed as equally spaced linear interpolations (for four postures) and extrapolations (for two postures) of the average vowel tongue postures, as illustrated in Fig. 1, *Right*, for one particular subject (S8). The combined set of vowel tongue postures and intermediate tongue postures constituted the nine target tongue postures used during the tongue-positioning task.

The tongue-positioning task. Fig. 2 illustrates the design of each trial. At the start of each trial, subjects were instructed to start from a tongue position for the consonant /l/. The consonant /l/ was chosen because it does not substantially influence subsequent tongue positioning, while providing a relatively stable reference in terms of tongue elevation (close to the palate). The target tongue posture, which was shown as four points in a straight line (Fig. 2, *Bottom Left*), was then presented along with a real-time display of the subject's sensor positions using the distorted spatial representation described above (black lines and circles in Fig. 2). Subjects were given 5 s after a visually displayed "go" signal to get as close as they could to the displayed target. The nine target tongue postures were displayed 10 times each in randomized order across trials. Each subject thus completed

90 trials (9 targets × 10 repetitions) in this tongue-positioning task. Note that precisely reaching the targets was not crucial since the main purpose of this task was to generate a range of tongue configurations that uniformly covered the full workspace, defined by the nine target tongue postures. This task was performed while holding the same bite block as during the production task.

The whispering task. Following each tongue-positioning trial, subjects were instructed to hold this reached tongue posture while whispering for a few seconds, with a reference countdown (around 2 s) being displayed on the computer screen. To help subjects maintain their position, the reached tongue posture was displayed on the screen. Auditory feedback was masked with white noise (80 dB SPL) presented through in-ear headphones. In order for the somatosensory subjects to keep their whispering intensity low, a level meter was displayed on the screen (Fig. 2). We carefully checked the possibility that despite the masking noise, whispers could provide auditory information. By evaluating the acoustic power of the whispers, both in the airborne and bone-conducted signals, and comparing these values with the power of the masking noise, we feel confident in rejecting this possibility (*SI Appendix*). This observation is in agreement with reports from subjects, when they were asked whether or not they could hear their whispers.

Since the whispering task lasted for a few seconds, some small variation in the tongue posture was expected during each whisper, as it is generally the case for any kind of postural control. In order to select a reached tongue posture that was representative of the whispering task, we considered for each trial the tongue configuration that was the closest to the median tongue posture calculated within the 1-s time window of best articulatory stability.

The somatosensory identification task. In this task we addressed the primary question of the study: Are subjects able to correctly identify vowels based on somatosensory information? On each trial, immediately after the whispering task, subjects were instructed to classify the reached tongue posture as one of the three vowels of interest, by selecting "é," "èè," or "a" on a screen (corresponding respectively to /e/, /ɛ/, and /a/ in French; Fig. 2, *Bottom Right*).

The auditory identification task. This part of the study involved a phonetic evaluation of the reached tongue postures based on auditory classification of the whispered speech. This task was performed by eight subjects (auditory subjects), who were not involved in any of the other tasks. Auditory subjects listened to the whispers one time each through headphones. The whispers of the somatosensory subjects were presented in separate blocks. The order of the whispers within each block was randomized and the order of presentation of blocks was balanced across auditory subjects. Auditory subjects labeled each sound by clicking on the appropriate button "a," "é," or "èè" on the screen. Each individual label provided by an auditory subject is called an auditory answer. For each whispered sound we extracted the most frequent vowel label among /e/, /ɛ/, and /a/ assigned to the auditory answers. We considered this label to be the canonical auditory label for this sound. Note that some whispered sounds could be associated with two different most frequent vowel labels; in that case, we assigned randomly the auditory label to one of these two vowel labels. For the sake of clarity, we call auditory answer the individual answer provided by auditory subjects and auditory label the most frequent answer among the eight auditory answers for a given whisper.

Data recording, data processing, and data analysis. Details about data recording, data processing, and data analysis (computation of the entropy of auditory answers, computation of the clustering scores using silhouette analysis, design of the Bayesian classifier, and construction of the identification curves) are provided in *SI Appendix*.

All data discussed in the paper and all programs used for their analysis will be made available to readers upon request.

ACKNOWLEDGMENTS. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2 013 grant agreement 339152, "Speech Unit(e)s"; principal investigator, Jean-Luc Schwartz), from the European Union's Horizon 2020 research and innovation program under the Marie Skłokodowska-Curie grant agreement 754490 (MultIscale precision therapies for NeuroDEvelopmental Disorders [MINDED] Program), and from the National Institute on Deafness and Other Communication Disorders grant R01DC017439. It was also supported by NeuroCoG "IDEX Université Grenoble Alpes: université de l'innovation" in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

1. J. A. Tourville, F. H. Guenther, The diva model: A neural theory of speech acquisition and production. *Lang. Cognit. Process.* **26**, 952–981 (2011).
2. B. Parrell, V. Ramanarayanan, S. Nagarajan, J. Houde, The facts model of speech motor control: Fusing state estimation and task-based control. *PLoS Comput. Biol.* **15**, e1007321 (2019).
3. G. Hickok, Computational neuroanatomy of speech production. *Nat. Rev. Neurosci.* **13**, 135–145 (2012).
4. J.-F. Patri, P. Perrier, J.-L. Schwartz, J. Diard, What drives the perceptual change resulting from speech motor adaptation? Evaluation of hypotheses in a bayesian modeling framework. *PLoS Comput. Biol.* **14**, e1005942 (2018).
5. B. J. Kröger, J. Kannampuzha, C. Neuschaefer-Rube, Towards a neurocomputational model of speech production and perception. *Speech Commun.* **51**, 793–809 (2009).
6. T. Gay, B. Lindblom, J. Lubker, Production of bite-block vowels: Acoustic equivalence by selective compensation. *J. Acoust. Soc. Am.* **69**, 802–810 (1981).
7. C. Savariaux, P. Perrier, J. P. Orliaguet, Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *J. Acoust. Soc. Am.* **98**, 2428–2442 (1995).
8. S. Tremblay, D. M. Shiller, D. J. Ostry, Somatosensory basis of speech production. *Nature* **423**, 866–869 (2003).
9. S. M. Nasir, D. J. Ostry, Somatosensory precision in speech production. *Curr. Biol.* **16**, 1918–1923 (2006).
10. J. S. Perkell *et al.*, A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *J. Phonetics*, **28**, 233–272 (2000).
11. A. M. Liberman, K. S. Harris, H. S. Hoffman, B. C. Griffith, The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* **54**, 358–368 (1957).
12. D. B. Pisoni, On the perception of speech sounds as biologically significant signals. *Brain Behav. Evol.* **16**, 330–350 (1979).
13. E. F. Chang *et al.*, Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* **13**, 1428–1432 (2010).
14. S. J. Lederman, R. L. Klatzky, Hand movements: A window into haptic object recognition. *Cognit. Psychol.* **19**, 342–368 (1987).
15. J. Robert-Ribes, J.-L. Schwartz, T. Lallouache, P. Escudier, Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise. *J. Acoust. Soc. Am.* **103**, 3677–3689 (1998).
16. P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
17. M. D. Ernst *et al.*, Permutation methods: A basis for exact inference. *Stat. Sci.* **19**, 676–685 (2004).
18. T. Ito, M. Tiede, D. J. Ostry, Somatosensory function in speech perception. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1245–1248 (2009).
19. C. A. Fowler, D. J. Dekle, Listening with eye and hand: Cross-modal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* **17**, 816–828 (1991).
20. B. Gick, D. Derrick, Aero-tactile integration in speech perception. *Nature* **462**, 502–504 (2009).
21. L. E. Bernstein, M. E. Demorest, D. C. Coulter, M. P. O'Connell, Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. *J. Acoust. Soc. Am.* **90**, 2971–2984 (1991).
22. P. A. Keating, "The window model of coarticulation: Articulatory evidence" in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, J. Kingston, M. E. Beckman, Eds. (Cambridge University Press, Cambridge, UK, 1990), chap. 26, pp. 451–470.
23. F. H. Guenther, S. S. Ghosh, J. A. Tourville, Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.*, **96**, 280–301 (2006).
24. X. Tian, D. Poeppel, The effect of imagination on stimulation: The functional specificity of efference copies in speech processing. *J. Cognit. Neurosci.* **25**, 1020–1036 (2013).
25. X. Tian, D. Poeppel, Dynamics of self-monitoring and error detection in speech production: Evidence from mental imagery and meg. *J. Cogn. Neurosci.* **27**, 352–364 (2015).
26. J. F. Houde, S. S. Nagarajan, Speech production as state feedback control. *Front. Hum. Neurosci.* **5**, 82 (2011).
27. D. M. Patri, R. C. Miall, M. Kawato, Internal models in the cerebellum. *Trends Cogn. Sci.* **2**, 338–347 (1998).
28. D. Recasens, A. Espinosa, Articulatory, positional and coarticulatory characteristics for clear/l/and dark/l: Evidence from two Catalan dialects. *J. Int. Phonetic Assoc.* **35**, 1–25 (2005).
29. P. Ladefoged, I. Maddieson, *The Sounds of the World's Languages* (Blackwell Publishers Ltd, Oxford, United Kingdom, 1996).