Multimodal Technologies and Interaction

MDPI

Article

Externalizing Tongue Articulation Through Forearm Touch: A Wizard-of-Oz Study of Tactile Articulatory Encoding for Arabic Letters

Abdullah Saleh Alshehri [1,*] and Paul Newbury [2]

[1] Department of Computer Science and Information Technology, Jubail Industrial College, P.O. Box 10099, Jubail Industrial City 31961, Saudi Arabia
[2] Department of Informatics, University of Sussex, Falmer, Brighton BN1 9RH, UK; p.newbury@sussex.ac.uk
* Correspondence: ushehri@gmail.com

Abstract

Speech articulation involves tongue movements that are largely hidden from direct perception, which makes them difficult to explain, observe, and guide during pronunciation instruction. This paper asks whether articulatory information can be externalized through touch and decoded by a listener. We introduce a tactile encoding scheme that maps regions of the tongue onto spatial locations on the forearm, combined with four tactile patterns (Tap, Hold, Trill, and Push) representing articulatory manner. To evaluate whether this encoding is interpretable, we conducted a Wizard-of-Oz study with six native Arabic speakers, in which cues were delivered manually by an experimenter rather than by a haptic device, so that the mapping could be examined before committing to a particular hardware implementation. Seven Arabic letters were used as targets, drawn from the full inventory of twenty-eight; participants were not informed that the target set was restricted, so responses were made against an open set. Following instruction on the general tactile–articulatory encoding principle, but before target-specific corrective feedback, participants identified the intended letter on 29.76% of trials, exceeding the nominal uniform-response reference levels defined for the open response set. After a brief exposure phase, accuracy reached 75.00% and response time fell from 14.48 s to 4.40 s. Critically, errors were not randomly distributed: they clustered on letters that are articulatorily adjacent to the target, and in several cases on letters that were never presented as targets (e.g., dāl, lām and nūn), a pattern that is difficult to explain by rote association alone. Some confusions involved targets encoded at neighboring forearm sites, suggesting possible spatial-discrimination limitations. We report these findings as preliminary evidence that tongue articulation may be represented in a decodable form through spatial touch. We make no claims regarding pronunciation learning, retention, or pedagogical efficacy, which this design cannot address.

Keywords: haptic feedback; tactile encoding; Arabic articulation; Wizard-of-Oz; multimodal interaction; cross-modal mapping

Check for updates

Academic Editor: Alexey Karpov

Received: 29 August 2026
Revised: 4 October 2026
Accepted: 3 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Achieving intelligible pronunciation requires precise control of articulatory movements that are often difficult to perceive, monitor, and guide during speech production [1,2]. Despite its critical role in communication, pronunciation has traditionally received less emphasis in language learning [1,3]. This imbalance is problematic, as adult learners frequently show high variability in pronunciation attainment [4]. These inaccuracies may

reflect perceptual and instructional challenges that are not fully addressed through auditory input alone [1,5]. A central obstacle to accurate production lies in the complexity of articulatory control, particularly the coordination of the tongue [2,6]. Because these articulatory movements occur behind the teeth, they are largely invisible to learners and difficult for instructors to demonstrate or evaluate, making articulatory state difficult to access directly through perception. Adult learners often lack conscious awareness of their own tongue gestures and may struggle to perform movements not acquired during first-language development [2,7].

To address this limitation, visual articulatory feedback has been used in pronunciation training to make otherwise hidden tongue movements available to learners [2,8]. Ouni [2] used real-time ultrasound imaging to provide visual feedback of tongue movement, while Katz and Mehta [8] used a real-time 3D tongue visualization system based on electromagnetic articulography (EMA) for novel speech-sound learning. These approaches demonstrate the potential value of making tongue articulation perceptible through external feedback. Alternative sensory modalities, including tactile feedback, provide another means of conveying articulatory or speech information [6,9–11].

Prior work has demonstrated that tactile systems can support both speech communication and pronunciation feedback: systems such as the Phoneme-Based Tactile Display and TAPS convey speech information through vibrotactile patterns, while HAPOVER and Touch to Speak apply tactile feedback to pronunciation support [9–13]. TAPS, in particular, incorporates articulatory features into its tactile phonemic code, including the use of forearm location to represent place of articulation [13]. The present work differs in its purpose and evaluation: rather than using tactile patterns primarily for speech reception, it externalizes tongue configuration onto the forearm as a production-oriented representation and examines whether this mapping can be interpreted before specific cue-to-letter associations are taught.

Answering it requires a design that separates two things that are easily conflated: whether a tactile encoding is interpretable at all, and whether it improves pronunciation. The second question can only be meaningfully investigated if the first is established: before testing whether tactile guidance improves pronunciation, the encoding must first be shown to convey interpretable articulatory information. If a mapping cannot be decoded by speakers who already command the target articulations, there is no basis for expecting it to guide speakers who do not. We therefore use native Arabic speakers as a methodological baseline. This choice is not a proxy for a learner population and does not license inferences about learners; rather, it provides a baseline in which the target articulations are already part of the participants' established speech repertoire, reducing the influence of incomplete L2 phonological knowledge.

Two features of the design are central to interpreting the results. First, cues were delivered manually by an experimenter rather than by a haptic device, following the Wizard-of-Oz paradigm [14,15], in which a human operator simulates system behavior so that a concept can be evaluated before hardware commits the design to a fixed set of actuation constraints. Second, although only seven Arabic letters served as targets, participants were not told this. Responses were therefore made against the full inventory of twenty-eight letters, which permits performance to be compared against nominal uniform-response reference baselines and constrains the extent to which accuracy can be attributed to a memorized correspondence between cues and a small, closed set.

The following research questions guide the investigation:

1. To what extent can tactile cues representing articulatory configurations be decoded into the intended targets after instruction on the general encoding principle but before

target-specific corrective feedback, relative to the nominal uniform-response reference levels defined for the open response set?
2. How do decoding accuracy and response time change after brief exposure to the mapping?
3. Does the structure of decoding errors follow articulatory and spatial adjacency, as would be expected if participants were interpreting the mapping, or is it unsystematic, as would be expected under rote association?

This paper makes three contributions. First, it introduces a tactile encoding scheme that maps the articulatory space of the tongue onto spatial locations on the forearm, together with a vocabulary of four patterns representing articulatory manner. Second, it provides evidence, from an open-set identification task, that this encoding can be decoded above the nominal uniform-response reference levels after instruction on the general encoding principle but before target-specific corrective feedback. Third, through an analysis of error structure, it suggests possible spatial-discrimination limitations, particularly for targets using the same tactile pattern at neighboring forearm locations, which points to a concrete design target for subsequent hardware implementations.

2. Related Work

To address the invisibility of articulators, Computer-Assisted Pronunciation Training (CAPT) has evolved beyond purely auditory models to incorporate visual and instrumental feedback systems [2,8,16]. Modern systems also use visual articulatory displays to illustrate hidden tongue gestures [2,8]. Additionally, instrumental techniques such as electropalatography (EPG), electromagnetic articulography (EMA), and ultrasound imaging have been utilized to provide real-time objective data of a learner's own articulation [2,8]. While these tools provide valuable articulatory information, they possess significant limitations for supporting speech production. Although 3D articulatory models can provide detailed visual feedback, mapping visual representations onto one's own articulatory motor system can be cognitively demanding [8]. Furthermore, they are either invasive, such as EMA, or produce noisy data, such as ultrasound images, that require expert interpretation.

While Automatic Speech Recognition (ASR) provides immediate diagnostic feedback, it typically indicates that an error occurred without specifying its underlying articulatory cause [17]. The pedagogical efficacy of these interventions may also depend on how learners process and integrate feedback, as explicit and implicit learning abilities can contribute differently to L2 pronunciation development [4]. Even high-quality visual data may create a cognitive bottleneck, as learners must monitor their own invisible articulatory movements [6]. This challenge is supported by prior research showing that an emphasis on visual accuracy may overlook functional intelligibility, defined as the ability to be understood by listeners rather than achieving a native-like accent [1,3].

Recent work in HCI and language learning has explored tactile and embodied feedback as a means of making otherwise intangible pronunciation phenomena more accessible. Grounded in embodied cognition, these approaches suggest that speech production is intrinsically linked to bodily movement. Existing research has demonstrated the feasibility of tactile speech and pronunciation systems, using wearable devices like the TActile Phonemic Sleeve (TAPS), HAPOVER, and Phoneme-Based Tactile Display, to provide speech-related information through vibrotactile patterns on the skin [9,10,13]. More broadly, haptic technology has been identified as a potentially valuable learning modality when integrated with appropriate pedagogical design [18]. In L2 pronunciation, individual corrective feedback has been shown to improve short-term comprehensibility compared with listening-only practice [19]. Real-time visual feedback of tongue movement has also been shown to support rapid gains in novel speech-sound production [8]. Beyond encoding design, reliable tactile

presentation also depends on consistent control of stimulus timing, intensity, and contact conditions. Device-based tactile systems can reduce such variability through calibrated and repeatable stimulus delivery.

The encoding examined here differs primarily in its purpose and organization. It uses a direct tongue-to-forearm spatial correspondence to externalize articulatory configuration as a production-oriented representation rather than primarily as a tactile speech-reception code. Clinical intraoral systems have provided error feedback for specific tongue placements, such as contact with the alveolar ridge, but operate within constrained clinical contexts and do not generalize to a full articulatory inventory [6]. Research on haptic relocation away from the fingertip has established the forearm as a viable site for spatially structured tactile information [20], while work on vibrotactile localization on the arm documents the resolution limits that any such scheme must respect [21]. Whether a production-oriented tongue-to-forearm mapping can be interpreted before specific cue-to-letter associations are taught has not, to our knowledge, been tested directly.

3. Materials and Methods

To answer the research questions, this study employed an exploratory proof-of-concept design to examine whether a tactile encoding of articulatory configurations could be decoded, how decoding performance changed after brief exposure, and how participants experienced the task. The study was conducted at the University of Sussex in Brighton, Southeast England, United Kingdom. Participants were recruited from the university community. This study was approved by the University of Sussex Ethics Committee (Approval No. 1380). All participants provided informed consent prior to participation. The methodology consisted of a controlled tactile-to-letter identification task in which tactile cues were mapped to articulatory tongue positions, followed by an exposure phase with corrective feedback and a post-test evaluation.

3.1. Wizard-of-Oz Cue Delivery

Tactile cues were delivered manually by an experimenter using finger contact, with no haptic device involved. This follows the Wizard-of-Oz paradigm, in which a human operator simulates the behavior of a system so that the interaction concept can be evaluated before implementation constrains it [14,15]. The paradigm is adopted here for a specific reason. Any haptic device imposes a fixed actuator layout, amplitude range, and temporal resolution, and a negative result obtained with such a device is ambiguous: it may indicate that the encoding is uninterpretable, or merely that the chosen hardware renders it poorly. Manual delivery reduces this ambiguity, permitting the articulatory mapping to be examined before hardware constraints are fixed and allowing the resulting requirements, particularly spatial resolution, to inform subsequent hardware design rather than being dictated by it.

The cost of this choice is that cue delivery was not mechanically uniform. Timing and contact force varied within the tolerances of manual execution, and this variability is a genuine limitation of the present data. To constrain manual variability, the six stimulation locations were marked on the forearm, all cues were delivered by the same experimenter using finger contact, and participants rested the selected forearm on the table in a consistent position. The six stimulation locations followed the same relative layout along each participant's forearm and were marked before testing. Tap was delivered as a single brief contact. Hold, Trill, and Push cues were delivered over approximately two seconds. Trill consisted of five quick successive contacts at the cued location. Hold consisted of sustained contact combined with a manually produced low-frequency oscillation of the fingertip; its frequency was not measured. Push, used for ḍād, was delivered by placing the thumb on location 1 and both index fingers on the two location-2 sites, then moving both index

fingers with slight pressure toward the thumb. The displacement of this movement was not measured. Absolute inter-site distances and contact-force values were not recorded. All participants received the same set of targets under the same procedure. A device-based replication is a necessary next step and is not substituted for by these results.

3.2. Design and Materials

A within-subjects design was used, in which all participants completed the identification task under tactile guidance. Targets were selected prior to the experiment on the basis of their articulatory relevance and reliance on tongue positioning, each corresponding to a specific configuration of the tongue. To support the encoding, the tongue was conceptualized as a set of discrete articulatory regions. Figure 1 presents a simplified top-view schematic showing the numbered articulatory locations used throughout the study. These locations were used consistently to describe articulatory targets during the task. The design employed a cross-modal mapping in which articulatory locations of the tongue were represented through spatial tactile cues delivered on the forearm, as shown in Figure 2. The forearm was selected for its accessibility and its relatively large surface area for spatial mapping [20].

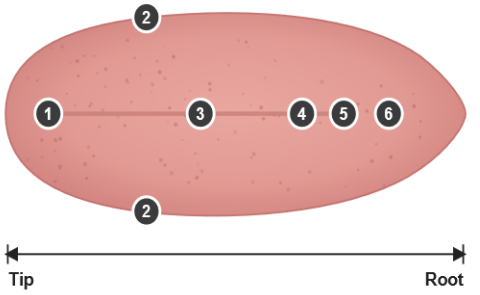


Figure 1. Conceptual top-view schematic of the tongue showing the numbered articulatory regions used for tactile articulatory guidance. The figure is intended as a schematic representation rather than an anatomically detailed model.



Figure 2. Forearm locations used for tactile articulatory guidance, with marked points corresponding to conceptual tongue regions in the spatial mapping.

The selected pronunciation targets and their corresponding representations are summarized in Table 1. Arabic script is presented alongside International Phonetic Alphabet (IPA) representations to facilitate international readability and to clarify the intended articulatory targets. In this study, "letter" refers to the categorical Arabic target used for participant responses, whereas the IPA notation represents the associated consonantal speech sound and articulatory target.

The seven targets were purposefully selected for this proof-of-concept study to represent articulatory configurations distributed across the tongue regions included in the proposed spatial mapping, from anterior to posterior locations, while also incorporating the tactile patterns used in the encoding scheme. The aim was therefore not to obtain a representative sample of all twenty-eight Arabic letters, but to test whether tongue-based articulatory information spanning the proposed mapping could be conveyed through forearm touch. Arabic sounds whose primary articulation depends mainly on structures outside the tongue-focused mapping were outside the scope of the present study.

Table 1. Selected Arabic pronunciation targets.

Tongue Location	Letter	IPA	Tactile Pattern
1	tā’	/t/	Tap (●)
5	khā’	/x/	Hold (vibration)
1	rā’	/r/	Trill (~~)
1 + 2	ḍād	/dˤ/	Push (→)
6	ghayn	/ɣ/	Hold (vibration)
4	qāf	/q/	Tap (●)
3	kāf	/k/	Tap (●)

In addition to spatial location, tactile articulatory guidance was further differentiated through a small set of distinct tactile patterns. Each pattern was designed to convey simplified articulatory feature information. These patterns were introduced to participants prior to the experimental task and were applied consistently throughout the study.

The tactile patterns were defined as follows:

- Tap (●): a brief, discrete tactile contact representing momentary articulatory closure, characteristic of plosive consonants.
- Hold (vibration): a sustained tactile contact representing prolonged constriction and airflow, corresponding to fricative or continuous sounds.
- Trill (~~): a rapid sequence of repeated tactile contacts representing oscillatory articulatory movement.
- Push (→): a directional tactile cue representing guided articulatory movement or emphasis.

Together, these four patterns were designed to represent the principal articulatory actions needed to distinguish the seven targets included in this proof-of-concept study, including closure, sustained constriction, oscillatory movement, and directional movement. They are not intended to constitute a complete representation of articulatory patterns across the Arabic sound system. During the experiment, tactile patterns were administered manually via finger contact at predefined forearm locations. A Tap was delivered as a brief touch-and-release at a single location. A Hold consisted of sustained contact with a manually produced low-frequency oscillation at one location. A Trill consisted of multiple rapid, perceptually distinct touches at the same location. A Push was delivered by maintaining contact while moving the fingers toward another mapped location, creating a directional tactile cue. All patterns were delivered according to the same predefined protocol across trials.

These general encoding principles were introduced to participants before the experimental task. No electronic haptic device was used in this study. Manual tactile delivery allowed the conceptual cue patterns to be evaluated without the constraints of a fixed actuator layout or hardware-specific actuation characteristics. This approach was intentionally adopted to examine the tactile–articulatory mapping before committing to device-specific constraints, although manual delivery introduces its own variability (Section 3.1).

3.3. Response Set and Chance Baseline

Seven letters served as targets, drawn from the twenty-eight letters of the Arabic alphabet. Participants were not informed that the target set was restricted, and no candidate set was displayed or described at any point. Participants were instructed to produce the Arabic letter that best corresponded to the articulatory configuration they perceived, with no constraint on which letter they could produce. Responses were therefore made against an open set.

These response spaces provide nominal uniform-response reference baselines rather than empirical estimates of participants' true guessing probabilities. If responses were uniformly distributed across all twenty-eight Arabic letters, expected accuracy would be 3.57%. An alternative nominal reference of 5.00% results from assuming uniform responses across approximately twenty letters involving lingual articulation. However, because participants knew that the task concerned tongue articulation, actual response probabilities may have been non-uniform. Accordingly, these values are used as descriptive reference points and should not be interpreted as definitive behavioral chance levels.

3.4. Participants

A total of six male native Arabic speakers (age range 26–38 years) participated in the study. All participants reported Arabic as their first language and indicated no history of speech, hearing, or neurological disorders that might affect speech production. Participants were recruited voluntarily from the local university community. Data collection was conducted from 22 April to 22 May 2026. Native Arabic speakers were selected on the basis of an established methodological principle in pronunciation technology research: before a novel feedback modality can be meaningfully evaluated for its pedagogical efficacy with L2 learners, it must first be shown to carry interpretable articulatory information under conditions where phonological knowledge is not a confounding variable. Using native speakers as participants provides a methodological baseline in which the target articulations are already part of the participants' established speech repertoire, thereby reducing but not eliminating the possibility that decoding difficulties arise from incomplete knowledge of the target sounds themselves. This allows the study to focus more directly on the interpretability of the tactile–articulatory mapping, while recognizing that native speakers may still differ in their explicit awareness of articulatory gestures.

3.5. Procedure

Each participant completed a tactile-to-letter identification task consisting of three phases: pre-test, training, and post-test. The task included a fixed set of seven Arabic pronunciation targets, each presented twice per phase, resulting in 14 trials in both the pre-test and post-test. Trial order within each phase was randomized. During the corrective-feedback exposure phase, the seven targets were presented in randomized order for up to three full rounds. After each tactile cue, participants received binary feedback indicating whether their response was correct or incorrect. Following an incorrect response, the same cue could be attempted up to three times before proceeding. At the end of each full round, participants who felt confident could proceed to the post-test; otherwise, they completed another randomized round, up to a maximum of three rounds. Participant 2 completed three corrective-feedback rounds; all other participants completed two.

At the beginning of the session, participants were introduced to the concept of tactile articulatory guidance. The experimenter explained the mapping between forearm locations and tongue positions, and participants were shown a schematic representation of tongue regions. Participants were introduced only to the general encoding principle; no specific tactile cue-to-letter associations were disclosed before the pre-test. Participants were then

asked to interpret each tactile cue and produce the closest corresponding Arabic letter based on their perceived articulatory configuration. Before testing, participants were asked which hand they preferred or mainly used, and tactile stimulation was delivered to the corresponding forearm; this choice was recorded for each participant. Figure 2 shows a right forearm for illustration only.

Tactile cues were delivered manually by the same experimenter using finger contact at one of six predefined forearm locations corresponding to conceptual tongue regions. Each cue combined a spatial location with a tactile pattern (e.g., Tap, Hold, Trill, or Push), applied consistently according to the predefined mapping. All participants received the same set of targets and underwent the same procedure.

For each trial, response time was measured from the onset of the tactile cue to the onset of the participant's spoken response. This procedure was designed to assess immediate interpretability and short-term adaptation to the tactile–articulatory mapping rather than long-term learning or retention.

3.6. Scoring

Responses were scored as correct or incorrect according to whether the participant identified the intended target, either by producing the corresponding consonantal sound or by naming the Arabic letter. Both response forms were treated as the same categorical target identity; pronunciation quality was not graded. Scoring was performed by the experimenter, a native speaker of Arabic with expertise in Arabic phonology.

This is a limitation of the present study, and it is the most serious one. The same individual delivered the cues and scored the responses, which admits expectancy effects and precludes blinded assessment. The categorical nature of the judgment constrains the scope for bias relative to a graded assessment of pronunciation quality but does not eliminate the possibility of expectancy bias. Independent blinded scoring should therefore be incorporated in future replications of the study.

Following the task, participants completed a short questionnaire on perceived demand and experience, rated on five-point Likert scales. The questionnaire comprised nine dimensions: mental demand, physical demand, temporal demand, perceived performance, effort, frustration, enjoyment, engagement, and perceived applicability. These responses are reported as descriptive observations only.

4. Results

4.1. Identification Accuracy Relative to Chance

Before any exposure to corrective feedback, participants identified the intended target on 29.76% of pre-test trials. The observed pre-test accuracy exceeded both nominal uniform-response reference levels: 3.57% for the full twenty-eight-letter inventory and 5.00% for the approximate twenty-letter lingual subset. These comparisons are descriptive because the participants' actual response probabilities cannot be assumed to have been uniform. This pre-test figure is the cleanest evidence the design provides, because at this point participants had received no feedback and had no opportunity to infer that the target set was restricted.

Following the exposure phase, mean accuracy rose to 75.00% and mean response time fell from 14.48 s to 4.40 s. One post-test kāf response was not recorded and was scored as incorrect, so post-test accuracy is calculated over all 84 trials (63 correct). Variability decreased on both measures (accuracy SD 15.92 to 9.85; response time SD 6.58 to 1.38). Table 2 summarizes both phases. All six participants showed higher post-test accuracy and lower response times than in the pre-test. An exploratory Wilcoxon signed-rank test yielded the same directional result for both measures ($W = 0$, $p = 0.031$); given the sample size this is reported as supportive rather than as a basis for inferential claims.

Table 2. Mean & SD of identification accuracy (%) and response time (s) in the pre-test and post-test phases.

Phase	Mean Acc. (%)	SD Acc.	Mean RT	SD RT
Pre-test	29.76	15.92	14.48	6.58
Post-test	75.00	9.85	4.40	1.38

The interpretation of the pre-post change requires care. Before the post-test began, participants had already encountered each of the seven targets twice during the pre-test and during two or three corrective-feedback exposure rounds (three for Participant 2, two for all others). They encountered each target twice more during the post-test. Participants may therefore have inferred that the target set was restricted, narrowing the effective response space and potentially inflating post-test accuracy independently of improved decoding. The post-test accuracy figure should therefore be interpreted cautiously because it may overestimate improvement attributable to decoding of the mapping itself. The pre-test provides the cleaner estimate of initial interpretability.

4.2. Error Structure

Error analysis further revealed systematic structure in participants' incorrect responses. Errors were classified as in-set when the response was another of the seven target letters and as out-of-set when it was any other Arabic letter; because participants were not informed that the target set was restricted, out-of-set responses remained available throughout both test phases. As shown in Table 3, out-of-set errors fell from 41 in the pre-test to 9 in the post-test, whereas in-set errors decreased less, from 18 to 11. The errors that remained after training therefore lay mostly within the mapped set and were concentrated among the posterior targets qāf, ghayn and kāf. Across both phases, the most frequent specific confusions were ḍād→lām (9 occurrences; 7 pre-test, 2 post-test), tā'→dāl (8; 7, 1), ghayn→qāf (6; 2, 4), qāf→kāf (5; 1, 4), tā'→nūn (4; 1, 3) and ghayn→khā' (4; 4, 0). Table 4 gives the complete response pattern for each target as counts out of 12 trials per phase; for example, qāf was identified as kāf in 1 of 12 pre-test and 4 of 12 post-test trials, whereas kāf was identified as qāf in 1 of 12 trials in each phase.

Table 3. Correct responses, in-set errors and out-of-set errors per target in the pre-test and post-test (12 trials per target per phase).

Target	Correct Pre	Correct Post	In-Set Errors Pre	In-Set Errors Post	Out-of-Set Errors Pre	Out-of-Set Errors Post
ت tā'	2	7	1	0	9	5
ك kāf	5	8 [a]	1	2	6	1
ق qāf	2	7	4	5	6	0
خ khā'	3	12	5	0	4	0
غ ghayn	2	8	7	4	3	0
ض ḍād	1	10	0	0	11	2
ر rā'	10	11	0	0	2	1
Total	25	63	18	11	41	9

[a] One post-test kāf response was not recorded (data-entry error) and was scored as incorrect; the kāf post-test row therefore sums to 11 and the post-test totals to 83 of 84 trials.

Table 4. Confusion matrix of target letters (rows) and responses (columns). Each cell shows pre-test/post-test counts out of 12 trials per target per phase; correct responses are on the diagonal in bold. Dāl, lām and nūn are out-of-set letters; Other pools all remaining out-of-set responses, including one unrecorded post-test kāf response. – = 0 in both phases.

Target	tā' ت	kāf ك	qāf ق	khā' خ	ghayn غ	ḍād ض	rā' ر	dāl د	lām ل	nūn ن	Other
tā' (ت)	**2/7**	–	–	–	–	–	1/0	7/1	–	1/3	1/1
kāf (ك)	–	**5/8**	1/1	0/1	–	–	–	1/0	1/0	1/0	3/2
qāf (ق)	–	1/4	**2/7**	1/0	2/1	–	–	–	–	1/0	5/0
khā' (خ)	–	2/0	3/0	**3/12**	–	–	–	–	–	–	4/0
ghayn (غ)	–	1/0	2/4	4/0	**2/8**	–	–	–	–	–	3/0
ḍād (ض)	–	–	–	–	–	**1/10**	–	–	7/2	2/0	2/0
rā' (ر)	–	–	–	–	–	–	**10/11**	–	–	1/1	1/0

First, three of the six most frequent confusions involved letters that were never presented as targets: dāl, lām and nūn. These out-of-set responses are difficult to explain by a simple closed-set cue-to-label retrieval account alone, since a participant guessing among memorized target labels would have no reason to produce them, and they are compatible with participants using articulatory information contained in the cues. For example, dāl shares place of articulation with tā' and differs primarily in voicing, and lām, like ḍād, may involve lateral tongue contact. Not every articulatorily close letter attracted responses, however: ṭā', the emphatic stop closest to ḍād, was given as a response to ḍād only once. These patterns therefore provide supportive, rather than conclusive, evidence that participants were interpreting articulatory information rather than relying only on memorized target labels.

Second, the qāf–kāf confusion suggests a possible spatial-discrimination limitation. Both targets used the Tap pattern but were mapped to neighboring forearm locations (4 and 3, respectively), and qāf→kāf was one of the few confusions that became more frequent after training, accounting for four of qāf's five post-test errors. Because pattern type was held constant for this pair, their confusion is consistent with difficulty distinguishing nearby stimulation sites. The ghayn–qāf confusion, which also increased at post-test, involved relatively nearby locations (6 and 4), but these targets differed in both location and tactile pattern, so this confusion cannot be attributed to spatial resolution alone. Overall, the observed confusion patterns suggest possible spatial-discrimination limitations of the mapping, particularly among the posterior targets, although the present design does not isolate spatial discriminability as the sole cause of error.

4.3. Participant Responses

Table 5 summarizes mean ratings across all nine questionnaire dimensions. Participants reported low physical demand ($M = 1.17$, $SD = 0.41$), temporal demand ($M = 1.50$, $SD = 0.55$), and frustration ($M = 1.00$, $SD = 0.00$), alongside high perceived performance ($M = 4.33$, $SD = 0.82$), enjoyment ($M = 4.50$, $SD = 0.55$), engagement ($M = 5.00$, $SD = 0.00$), and perceived applicability ($M = 4.83$, $SD = 0.41$). Mental demand was moderate ($M = 3.33$, $SD = 1.21$).

Table 5. Subjective ratings across dimensions.

Dimension	Mean Rating	SD
Mental demand	3.33	1.21
Physical demand	1.17	0.41
Temporal demand	1.50	0.55
Perceived performance	4.33	0.82
Effort	2.33	1.37
Frustration	1.00	0.00
Enjoyment	4.50	0.55
Engagement	5.00	0.00
Perceived applicability	4.83	0.41

These figures are reported for completeness and should be treated with caution. Several items show no variance across participants, and the questionnaire was administered by the experimenter who conducted the session, conditions under which agreement ratings are susceptible to demand characteristics. We draw no conclusions about cognitive load from these data and note only that no participant reported the task as burdensome or frustrating enough to impede completion.

5. Discussion

The present study explored tactile articulatory encoding as a means of representing otherwise hidden tongue configurations, focusing on immediate decoding performance and user experience rather than pronunciation learning outcomes. This section discusses the findings in relation to prior work on pronunciation instruction and technology-mediated feedback and considers their implications for future pronunciation support tools.

5.1. Interpretation of Decoding Performance

The central result is that a tactile representation of tongue articulation was decoded above the nominal uniform-response reference levels before letter-specific corrective feedback. Participants had been introduced to the general encoding principle but had not been taught specific cue-to-letter associations or received target-specific corrective feedback; nevertheless, they identified the intended target at a rate exceeding the nominal uniform-response reference levels. Because actual response probabilities may have been non-uniform, this comparison is interpreted as descriptive rather than as a definitive estimate of above-chance performance. These findings are consistent with the mapping conveying interpretable articulatory information following instructions in the general encoding principle.

The error structure supports the same conclusion from a different direction. Responses that fell outside the presented target inventory, and that stood in close articulatory relation to the intended letter are difficult to explain by learned cue-to-label associations alone. Out-of-set responses such as dāl, lām and nūn, none of which was presented as a target, are consistent with participants drawing on articulatory information rather than relying only on memorized cue-to-label associations. Several observed confusions also follow articulatory relations among letters. The most frequent, ḍād (ض) confused with lām (ل) in nine responses, links two letters whose articulation involves the sides of the tongue, while the confusions of tā’ (ت) with dāl (د) and with nūn (ن) involve letters that share the same tongue-tip place and differ mainly in voicing and nasality, features the encoding does not represent. This is consistent with participants decoding the tongue region conveyed by the cue but lacking information about features the encoding leaves out, although this account remains interpretive. However, without scrambled-mapping control, the present design

cannot determine the mechanism responsible for this performance or exclude contributions from explicit instruction, phonological reasoning, repetition, or other cues.

It should also be noted that four of the seven targets (khā', ghayn, qāf, and kāf) were assigned distinct spatial locations and could therefore, in principle, be identified from location information without relying on the tactile pattern. The present design does not isolate the independent contributions of spatial location and tactile pattern to decoding performance; this would require a design in which these encoding dimensions are systematically controlled or varied.

This is a claim about signal interpretability and nothing more. The study does not show that tactile cues improve pronunciation, that any effect persists, or that the approach benefits learners. Participants already possessed the articulations in question; the task asked them to recover a specification, not to acquire a capability. Establishing interpretability is a precondition for investigating pedagogical value, not evidence of it.

5.2. Spatial Discriminability as a Design Constraint

The distribution of errors suggests that spatial discriminability contributed to some of the observed confusions. The qāf–kāf confusion is particularly informative because both targets used the same Tap pattern at neighboring forearm locations 4 and 3, making difficulty in spatial discrimination a plausible contributor. The ghayn–qāf confusion, however, involved differences in both location and tactile pattern and therefore cannot be attributed to spatial resolution alone. Vibrotactile localization on the arm is influenced by stimulation-site location and inter-site separation [21], and the present pattern is consistent with that literature.

This has a concrete design implication. The six mapped locations were distributed along the forearm at spacings that manual delivery could produce but that may be close to the limits of spatial discrimination at that site. Improving performance may therefore depend partly on increasing the effective spatial separation of encoded regions, while other sources of confusion should also be considered; possible approaches include recruiting an additional dimension such as pattern or intensity to disambiguate neighboring places or relocating portions of the mapping to sites of higher acuity [20]. Each is testable, and each is a hardware question that this study was designed to inform.

5.3. Relation to Prior Tactile Speech Systems

Prior tactile speech systems have used tactile codes for both speech reception and pronunciation feedback, and some, particularly TAPS, incorporate articulatory features such as place of articulation into their encoding [10,11,13]. The scheme examined here differs primarily in its purpose and organization: it externalizes tongue configuration onto the forearm as a production-oriented representation and examines whether that mapping can be interpreted before specific cue-to-letter associations are taught. The trade-off is that a structured code inherits the confusability of the structure it represents, as the articulatorily related confusions reported in Section 4.2 illustrate. Whether structured encoding retains an advantage over arbitrary encoding once both are learned is an open question this study does not address.

5.4. Limitations

The most serious limitation concerns scoring. The experimenter who delivered the cues also judged the responses, so assessment was not blinded and expectancy effects cannot be excluded. Retrospective blinded rescoring was not possible for the present data. This potential source of bias should be considered when interpreting the reported accuracy

figures, and future replications should incorporate independent scoring by raters blinded to the presented cue.

Cue delivery was manual and therefore not mechanically uniform in timing or contact force. This is inherent to the Wizard-of-Oz paradigm and was accepted deliberately in exchange for the ability to examine the mapping before hardware constraints are fixed, but it means the results characterize the encoding rather than any deployable system.

The sample of six participants supports a feasibility assessment and does not support population-level inference. The seven targets constitute a partial sample of the Arabic inventory, chosen for articulatory coverage, and performance on unmapped letters is unknown. There was no control condition; a comparison against a spatially scrambled mapping would isolate the contribution of the articulatory structure specifically and would strengthen the interpretability claim considerably at modest cost.

There was no delayed test, and none of these findings speak to retention. The post-test accuracy figure is subject to the inference concern described in Section 4.1 and should be read as an upper bound. The questionnaire data are descriptive, collected under conditions favorable to demand characteristics, and support no conclusions about cognitive load.

Finally, all participants were native speakers of the target language. The results suggest that the encoding is decodable by speakers who already possess the articulations; whether it can guide speakers who do not is a separate empirical question requiring a separate study.

5.5. Future Work

Three lines follow directly. First, a blinded replication with independent scoring and a scrambled-mapping control would address the two design weaknesses that most constrain the present claims. Second, the possible spatial-discrimination limitations suggested in Section 5.2 should be characterized systematically, establishing the minimum separation at which encoded regions remain reliably distinguishable on the forearm, which would determine the actuator layout for device implementation. A device-based implementation should also provide repeatable control of stimulus timing and intensity/contact force, reducing the variability inherent in manual tactile delivery. Third, once the encoding is validated instrumentally, the question of pedagogical value becomes tractable: a study with second-language learners, a control condition, a delayed post-test, and transfer items not presented during training would test whether interpretability translates into articulatory benefit. This is particularly relevant for future testing with adult L2 learners, as previous work has shown that non-native learners may experience difficulty discriminating unfamiliar Arabic sound contrasts [22]. That study is the appropriate setting for claims about pronunciation instruction, and the present work is a precondition for it rather than a preliminary version of it.

6. Conclusions

This paper examined whether tongue articulation can be represented in a form recoverable through touch. Using a Wizard-of-Oz protocol with six native Arabic speakers and an open response set, pre-test decoding accuracy exceeded the nominal uniform-response reference levels before letter-specific corrective feedback, with post-test accuracy reaching 75% after brief exposure. Out-of-set errors that were articulatorily related to the intended targets provide supportive evidence that participants were using articulatory information rather than relying solely on cue-to-label associations. The prominent qāf–kāf confusion, involving the same tactile pattern at neighboring forearm locations, further suggests that spatial discriminability is an important consideration for subsequent hardware design. These findings provide preliminary evidence that articulatory geometry can be represented in a

decodable tactile form. Whether such encoding supports pronunciation learning remains untested and requires a study designed specifically for that question.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/mti10110116/s1. Table S1: De-identified trial-level data; Table S2: Questionnaire data; Table S3: Training rounds per participant; Table S4: Complete pre-test and post-test confusion matrices.

Author Contributions: Conceptualization, A.S.A.; methodology, A.S.A. and P.N.; investigation, A.S.A.; data curation, A.S.A.; formal analysis, A.S.A.; visualization, A.S.A.; writing—original draft preparation, A.S.A.; writing—review and editing, A.S.A. and P.N.; supervision, P.N.; project administration, A.S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved by the University of Sussex Faculty Research Ethics Committee (reference number 1380) on 15 April 2026.

Informed Consent Statement: Written informed consent was obtained from all participants.

Data Availability Statement: De-identified trial-level response data, questionnaire data, participant training rounds and the complete confusion matrices are provided in the Supplementary Material. Further information is available on reasonable request from the corresponding author.

Acknowledgments: We thank all participants for their time and engagement in the study. The authors used ChatGPT version 5.6 tool for language refinement during manuscript preparation. All intellectual content, analysis, interpretation, and final text were reviewed and verified by the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
CAPT	Computer-Assisted Pronunciation Training
ECA	Embodied Conversational Agent
EMA	Electromagnetic Articulography
EPG	Electropalatography
HCI	Human–Computer Interaction
IPA	International Phonetic Alphabet
L2	Second Language
TAPS	TActile Phonemic Sleeve

References

1. Setter, J.; Jenkins, J. State-of-the-Art Review Article. *Lang. Teach.* **2005**, *38*, 1–17. [CrossRef]
2. Ouni, S. Tongue control and its implication in pronunciation training. *Comput. Assist. Lang. Learn.* **2014**, *27*, 439–453. [CrossRef]
3. Derwing, T.M.; Munro, M.J. *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2015. [CrossRef]
4. Suzukida, Y. The Contribution of Individual Differences to L2 Pronunciation Learning: Insights from Research and Pedagogical Implications. *RELC J.* **2021**, *52*, 48–61. [CrossRef]
5. Smotrova, T. Making Pronunciation Visible: Gesture in Teaching Pronunciation. *Tesol Q.* **2017**, *51*, 59–89. [CrossRef]
6. Jiang, B.; Biyani, S.; Park, H. A Wearable Intraoral System for Speech Therapy using Real-Time Closed-Loop Artificial Sensory Feedback to the Tongue. In Proceedings of the 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER), San Francisco, CA, USA, 20–23 March 2019; pp. 118–121. [CrossRef]
7. El Haimeur, A. Navigating Unshared Sounds: Challenges and Strategies in English Speakers' Acquisition of Arabic Emphatic and Pharyngeal Phonemes. *J. Afr. Lang. Teach. Assoc.* **2025**, *12*, 54–72.
8. Katz, W.F.; Mehta, S. Visual Feedback of Tongue Movement for Novel Speech Sound Learning. *Front. Hum. Neurosci.* **2015**, *9*, 612. [CrossRef] [PubMed]

9. Ilhan, R.; Kaçanoğlu, K. HAPOVER: A Haptic Pronunciation Improver Device. *IEEJ Trans. Electr. Electron. Eng.* **2024**, *19*, 985–992. [CrossRef]
10. Reed, C.M.; Tan, H.Z.; Perez, Z.D.; Wilson, E.C.; Severgnini, F.M.; Jung, J.; Martinez, J.S.; Jiao, Y.; Israr, A.; Lau, F.; et al. A Phonemic-Based Tactile Display for Speech Communication. *IEEE Trans. Haptics* **2019**, *12*, 2–17. [CrossRef] [PubMed]
11. Sharon, A.; Yozevitch, R.; Holdengreber, E. Touch to Speak: Real-Time Tactile Pronunciation Feedback for Individuals with Speech and Hearing Impairments. *Technologies* **2025**, *13*, 345. [CrossRef]
12. Reed, C.M.; Tan, H.Z.; Jones, L.A. Haptic Communication of Language. *IEEE Trans. Haptics* **2023**, *16*, 134–153. [CrossRef] [PubMed]
13. Tan, H.Z.; Reed, C.M.; Jiao, Y.; Perez, Z.D.; Wilson, E.C.; Jung, J.; Martinez, J.S.; Severgnini, F.M. Acquisition of 500 English Words through a TActile Phonemic Sleeve (TAPS). *IEEE Trans. Haptics* **2020**, *13*, 745–760. [CrossRef] [PubMed]
14. Dahlbäck, N.; Jönsson, A.; Ahrenberg, L. Wizard of Oz studies: Why and how. *Knowl. Based Syst.* **1993**, *6*, 258–266. [CrossRef]
15. Kelley, J.F. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.* **1984**, *2*, 26–41. [CrossRef]
16. Thomson, R.I. Computer Assisted Pronunciation Training: Targeting Second Language Vowel Perception Improves Pronunciation. *Calico J.* **2011**, *28*, 744–765. [CrossRef]
17. Dai, Y.; Wu, Z. Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Comput. Assist. Lang. Learn.* **2023**, *36*, 861–884. [CrossRef]
18. Crandall, R.; Karadoğan, E. Designing Pedagogically Effective Haptic Systems for Learning: A Review. *Appl. Sci.* **2021**, *11*, 6245. [CrossRef]
19. Dlaska, A.; Krekeler, C. The short-term effects of individual corrective feedback on L2 pronunciation. *System* **2013**, *41*, 25–37. [CrossRef]
20. Sarac, M. Haptic Relocation Away from the Fingertip: Where, Why, and How. *IEEE Trans. Haptics* **2025**, *18*, 470–481. [CrossRef] [PubMed]
21. Cholewiak, R.W.; Collins, A.A. Vibrotactile localization on the arm: Effects of place, space, and age. *Percept. Psychophys.* **2003**, *65*, 1058–1077. [CrossRef] [PubMed]
22. Al Mahmoud, M.S. Discrimination of Arabic contrasts by American learners. *Stud. Second Lang. Learn. Teach.* **2013**, *3*, 261–292. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.