# STOP VOICING AND DEVOICING AS ARTICULATORY TASKS: A CROSS-LINGUISTIC RT-MRI STUDY

Yubin Zhang and Louis Goldstein
Department of Linguistics, University of Southern California

## ABSTRACT

The current study tests the hypothesis that voicing and devoicing of stop consonants are phonatory tasks that are accomplished by a variable synergy of laryngeal and supraglottal articulators. We investigate the cross-linguistic differences in that synergy using RT-MRI of connected speech. French voiced stops were observed to have larger oral aperture, shorter oral constriction duration, higher velum position, more advanced tongue root, and larger post-constriction cavity than voiceless stops. In English, fewer articulatory dimensions are involved. Voiced stops have more advanced tongue root, lower larynx, and larger post-constriction cavity than voiceless stops. Moreover, by comparing nasals with voiced and voiceless stops, we found that the supraglottal adjustments are primarily employed for the devoicing rather than the voicing task. The results are consistent with the voicing/devoicing goal hypothesis, but languages differ in the articulatory dimensions used for this goal.

**Keywords**: stop, voicing, articulation, French, English.

## 1. INTRODUCTION

The stop voicing categories have been described as contrasting in voicing and aspiration [1], [2]. Languages with a two-way voicing contrast, are classified into two types of systems— the 'true voicing' and 'aspirating' languages [1], [2]. The 'true voicing' languages, such as French, Spanish, Russian and Portuguese, have rigorous closure voicing for voiced stops whereas their voiceless stops lack closure voicing and post-closure aspiration. For 'aspirating' languages, such as English and Mandarin, the contrast is considered as primarily in aspiration. These languages all have a category called voiceless aspirated stops with no closure voicing but post-closure aspiration. The other category typically does not have closure voicing and can be described as phonetically voiceless unaspirated stops. Note that while English is described as an 'aspirating' language, the latter category is not the canonical voiceless unaspirated stops. In the literature, it is also called phonologically voiced stops or lax voiceless stops. Closure voicing is generally lacking in domain-initial positions and intervocalic closure voicing is not always robust [1], [3].

The articulatory basis for voicing distinction has been proposed to lie in oral-laryngeal timing in models like articulatory phonology [4], [5] (see Figure 1). For both voiceless unaspirated and aspirated stops, the glottal opening gesture is timed in-phase with oral closing gesture, resulting in the silent closure interval [6]. For voiceless unaspirated stops with short-lag voicing onset time (VOT), the glottal closing gesture occurs before the oral release gesture, whereas for voiceless aspirated stops with long-lag VOT, the onset of the glottal closing gesture occurs near oral release [6], [7]. For phonologically voiced stops, only the oral constriction gesture is supposed to be activated. The presence of vocal fold vibration in the closure interval is not assumed to be specifically controlled. It is regarded as the default glottal state once phonation is initiated and when an active glottal opening-and-closing gesture is absent.

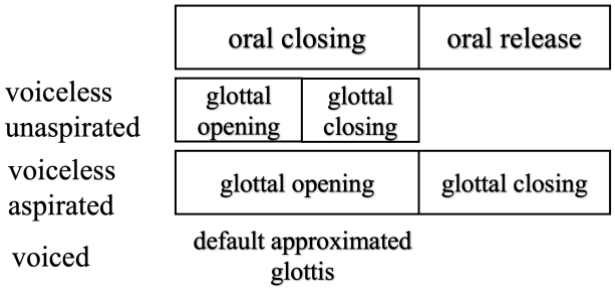


**Figure 1**: Illustration of the articulatory basis for voicing.

A potential problem with this proposal is that closure voicing does not always occur automatically with adducted vocal folds. Closure voicing is subject to the aerodynamic voicing constraint [8], [9]. With appropriately approximated and tensed vocal folds, a sufficient transglottal pressure difference, i.e., sufficiently larger subglottal pressure than intraoral pressure, is also needed to sustain continuous glottal airflow. However, the vocal tract constriction in stops quickly decreases the transglottal pressure. Without additional adjustments, voicing would cease quickly. Active glottal and supra-glottal articulations, such as vocal folds slackening [10], post-constriction cavity enlargement like larynx lowering and tongue root advancement [8], [11], oral/nasal leakage [12] and constriction duration reduction, [13] have been shown to be possible mechanisms for sustaining voicing [8], [12]. Terminating or inhibiting voicing can also require mechanisms in addition to the glottal opening gesture, like tightly adducting the vocal folds and stretching the vocal folds [7], [14], [15].

To account for these observations, some researchers suggest more global task goals, such as aerodynamic goals [16], [17]. [16] postulate a transglottal pressure task during voiced stop production. For devoicing in voiceless stops, they propose that the glottal opening gesture is the primary task involved. Nevertheless, the authors agree that other gestures like achieving a high F0 task goal by stretching and stiffening the vocal folds may also contribute to devoicing. Thus, [16] further discuss the possibility of postulating a 'voicing/devoicing' task that engages a flexible synergy comprising articulator components like glottal width, glottal tension, total lung force, and post-constriction cavity volume.

The current rt-MRI articulatory study investigates the cross-linguistic differences in the synergies comprising potential phonatory 'voicing/devoicing' tasks by examining the supra-glottal articulations during production of contrasting voiced and voiceless stops in French and English. For 'true voicing' languages like French, multiple articulator components are expected to be involved in achieving the voicing/devoicing tasks. However, for languages like English without rigorous closure voicing, fewer and less consistent articulatory dimensions might be employed in the voicing/devoicing tasks [12].

## 2. METHODS

The French articulatory data consist of mid-sagittal real-time MRI videos acquired from 5 male and 5 female French speakers aged 29 ± 8 years [18]. The speakers read aloud the speech materials, consisting of 77 manually constructed sentences with an almost-exhaustive coverage of the French speech sounds in various phonetic contexts (see [18] for more details). The MRI images were collected on a Siemens Prisma 3T scanner with a frame rate of 50 frames/s, image size of 136 × 136 and resolution of 1.6 mm × 1.6 mm. The audio was recorded in the MRI scanner with a sampling rate of 16kHz and then noise-cancelled using the algorithm described in [19]. The audio was automatically force-aligned with the transcriptions of words and phonemes using the French speech recognition system Astali. The English articulatory data were taken from the recordings of 7 participants in the publicly available USC-TIMIT real-time MRI database [20]. The participants were native speakers of American English (2 male, and 5 female, aged 30 ± 9 years). They read 460 sentences from MOCHA-TIMIT, a database designed to elicit all the English phonemes in various phonological and prosodic contexts. MRI data were acquired on a Signa Excite HD 1.5T scanner. The image resolution is 2.9 × 2.9 mm. The image size is 68 × 68. The reconstructed frame rate is 23.18 frames/s. The audio was simultaneously recorded at a sampling rate of 20kHz and then noise-cancelled using the method described in [22]. The audio has been force-aligned with text annotations of phonemes, words and sentences using SailAlign [21].

The articulatory analysis was based on stop and nasal tokens in the #_V, V_V, V_# contexts with oral vowels (# indicates utterance boundaries). Nasals were included as a neutral condition against which to evaluate articulator contributions to voicing and devoicing tasks. Velar nasals were not included because French does not have native velar nasals and most English velar nasals are from the 'ing' suffix. We analyzed a total of 1479 and 3910 stop tokens for French and English respectively.

The articulator contours in the MRI videos were tracked using the region-based segmentation method described in [22]. We first selected a reference frame during speaking without close approximation of articulators, and manually constructed an initial template locating the contours of the articulator segments such as tongue, hard palate, velum, lips, jaw, pharyngeal wall, epiglottis, etc. (see Figure 2 top left panel). Then, the initial template was registered to each frame of the MRI video and the air-tissue boundaries of the articulators were extracted using a hierarchical gradient descent procedure.

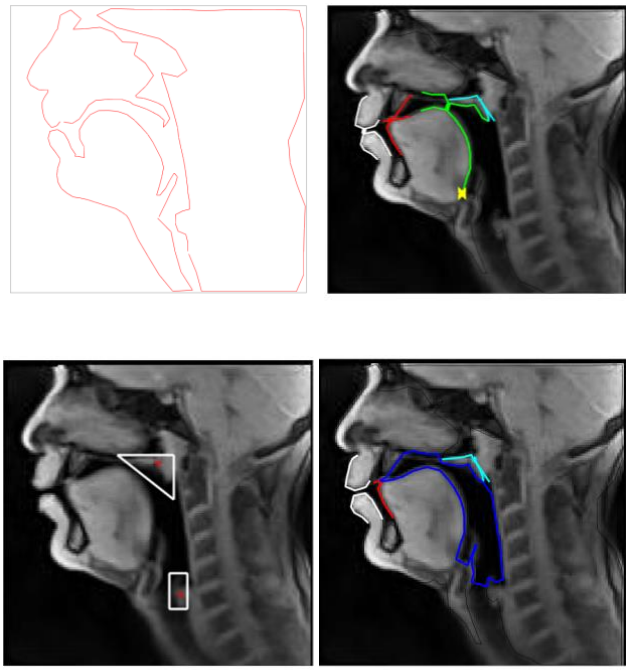

**Figure 2**: The region-based segmentation and centroid tracking analyses. Top left: the initial template; Top right: labial (white), alveolar (red), velar (green) and velic (cyan) search regions; The displayed frame is one where maximal alveolar constriction is achieved. The straight line denotes the minimal distance between two articulator contours. The yellow cross denotes the tongue root position; Bottom left:

the regions of interest (ROIs, triangle: velum; rectangle: larynx) and tracked centroids; Bottom right: post-constriction cavity mid-sagittal plane (blue).

Four search regions were identified to extract constriction aperture time series, i.e., the minimal distance between two contour lines (Figure 2 top right)— (1) lip constriction: between upper and lower lips, (2) alveolar constriction: between the tongue and alveolar ridge, (3) velar constriction: between the tongue and palate and (4) velic constriction: between the upper boundary of the velum and upper pharyngeal wall. The algorithm generates one single contour for the hard palate and tongue respectively. To separate the alveolar and velar search regions, we manually marked the boundary between the alveolar ridge and hard palate for each speaker by inspecting the reference image. The x-coordinate of the marked boundary point was used to divide the two regions. The generated velum contour was also divided into two parts: the upper and lower contours. The lower velum contour was used for the velar search region whereas the upper velum contour for the velic region. The upper pharyngeal wall was defined as the pharyngeal contour above the lowest velum point.

Horizontal tongue root movement was examined using the x-coordinate of tongue root position over time (Smaller x means more advanced). The tongue root point was defined as the point with the smallest y-coordinate on the tongue contour (the yellow cross in Figure 2 top right panel). Vertical larynx and velum movement trajectories were generated using a centroid-tracking method [23]. The algorithm automatically tracks the intensity-weighted centroid of an object in a user-defined region of interest (ROI). For each speaker, we manually specified a rectangular larynx ROI and a triangular velum ROI, which roughly cover the movement range for the arytenoid cartilage and velum respectively (see Figure 2, bottom left). The y-coordinates of the larynx and velum centroids over time were taken as the time series of vertical larynx and velum movements (larger y means higher position). The time series of mid-sagittal post-constriction cavity area was computed by combining all the articulator contours ranging from the constriction location (the straight line denoting minimal distance between oral articulators) to the middle of arytenoid cartilage (y-coordinate of the larynx centroid) at each time point (see the blue contour in Figure 2, bottom right). All the time series were smoothed with a span of 15 data points using the `rloess` function in `Matlab`.

The maximum constriction and gesture duration of oral gestures were computed using the `findgest` algorithm in `MVIEW` [24]. We searched for a velocity minimum closed to the mid-point of the force-aligned acoustic consonant interval. The aperture size at the velocity minimum was taken as the maximum constriction (MAXC). Then, two velocity peaks (PVEL1 and PVEL2) preceding and following MAXC were identified. Gestural duration was calculated as the difference between gestural onset and offset, which are defined using the 20% threshold criterion (onset: 20% of the range between the velocity minimum preceding PVEL1 and PVEL1; offset: 20% of the range between PVEL2 and following velocity minimum). Labelling errors were corrected manually. Non-oral-constriction articulatory parameters, i.e., velic aperture, velum height, horizontal tongue root position, larynx height, and post-constriction cavity area, were measured at the time point of MAXC.

We fit linear mixed-effects models with factors `voicing` and `place of articulation (POA)` for each articulatory dimension using `lmer()` in the `R` package `lme4` [25]. Their interaction was included if justified by the likelihood ratio test. Two separate series of models were fit—*v=vl* and *v=vl=nas* models. In *v=vl* models, `voicing` has two levels—*voiced* and *voiceless* with *voiceless* as the reference level. In *v=vl=nas* models, *nasal* was included as a reference level for the factor `voicing`. Only bilabial and alveolar data were analyzed because velar nasals were not included. For velum-related and cavity size measures, the nasal condition is no longer a neutral condition. Comparison with the nasal condition is not meaningful for these measures. Thus, only the *v=vl* model is reported in such cases.

| | **French** | **English** |
|---|---|---|
| oral aperture | v>vl: **;<br>vl<nas: **;<br>v-nas: n.s. | n.s. |
| oral constriction duration | v<vl: *;<br>vl-nas: n.s.;<br>v-nas: n.s. | v-vl: n.s.;<br>vl>nas: *;<br>v-nas: n.s. |
| velic aperture | v-vl: n.s. | v-vl: n.s. |
| velum height | v>vl: *; | n.s. |
| horizontal tongue root position | v<vl: *;<br>vl>nas: *;<br>v-nas: n.s. | v<vl: ***;<br>vl>nas: *;<br>v-nas: n.s. |
| larynx height | v-vl: n.s.;<br>v>nas: ***;<br>vl>nas: * | v<vl: ***;<br>v>nas: *** ;<br>vl>nas: *** |
| mid-sagittal post-constriction cavity area | v>vl: *** | v>vl: *** |

**Table 1**: The results for the `voicing` effects. Notations: v: voiced; vl: voiceless; nas: nasal; -: non-significant comparison; n.s.: non-significant, * $p \leqslant 0.05$, ** $p < 0.01$, *** $p < 0.001$. `POA` effects are not reported here.

## 3. RESULTS

The results for the `voicing` effects are shown in Table 1. The *v=vl* models reveal that French voiced stops have larger oral constriction aperture (v>vl), shorter oral constriction duration (v<vl), higher velum (v>vl), more advanced tongue root (v<vl), larger mid-sagittal post-constriction cavity area (v>vl) than voiceless stops. In English, voiced stops only have more advanced tongue root (v<vl), lower larynx (v<vl), and larger mid-sagittal post-constriction cavity area (v>vl) than voiceless stops. In the *v=vl=nas* models, when a significant comparison involving nasals was found, it is generally the nasal-voiceless comparison (e.g., vl<nas for oral aperture in French), except for larynx height (v, vl>nas). This result suggests that `voicing` effect is largely driven by the difference between nasals and voiceless stops.

## 4. DISCUSSION

The current results show that both languages engage multiple supra-glottal articulatory components in the voicing contrast, consistent with the hypothesis that voicing/devoicing can be more global linguistic motor tasks achieved by a synergy of laryngeal and supraglottal articulatory components [16], [17]. However, the exact task variable for voicing/devoicing remains to be investigated. [16] suggests voicing amplitude whereas [26] focuses more on durational measures like VOT. Moreover, a further way to gain insight on the synergy might be to investigate the compensatory or cooperative behaviors among the articulatory dimensions by more controlled experiments.

We also found that the difference in supraglottal articulations between voiced and voiceless stops is largely driven by the difference between nasals and voiceless stops. Previous research generally assumes that the supra-glottal articulations in stop voicing contrast are employed for sustaining voicing rather than devoicing. Terms like cavity expansion and oral/nasal leakage are frequently used [9], [13], [17]. However, our results suggest that in both languages, devoicing mechanisms, like aperture size reduction and tongue root retraction, seem to be involved more than voicing-sustaining mechanisms, at least for the dimensions where nasal data is meaningful. Most previous studies like [13], [26] did not include nasals as a neutral condition and could not evaluate the articulator contributions to voicing and devoicing. The English data from [9] included nasals for dimensions like tongue root. They found that for tongue root position, voiceless stops and nasals show more similarity than voiced stops and nasals, indicating more involvement of voicing-sustaining mechanisms. However, since they only recruited one speaker and provided no quantitative analysis, the discrepancy remains to be resolved by future studies. Note that nasals do not seem to be a valid neutral condition for larynx height. In both languages, stops have a higher larynx than nasals. This result is consistent with [9], but the reasons remain murky.

For language-specific differences, consistent with our prediction, and previous findings on cross-linguistic voicing/devoicing control [12], French speakers employ more supraglottal articulatory adjustments than English speakers. French speakers engage articulations like adjusting oral aperture size, oral constriction duration, velum height, horizontal tongue root position, and the post-constriction cavity size. English speakers exhibit differences in horizontal tongue root position, larynx height, and post-constriction cavity size, consistent with [8], [27]. Both languages use cavity expansion/shrinkage mechanisms, but the synergies show language-specific differences, i.e., expansion/shrinkage in the tongue root and velar regions for French but in the tongue root and larynx regions for English. Moreover, for aperture size measures, French speakers alter the oral aperture size, but neither French nor English speakers alter the velic aperture size. One explanation is that the effect size of velic leakage is extremely small [12]. The spatial resolution of the rt-MRI images does not allow us to reliably identify small nasal aperture adjustments. It is also likely that unlike the original proposal in articulatory phonology, other dimensions for velum control like velum height are more crucial than the velic constriction gesture.

One limitation of the current study is that voicing-related acoustic measures cannot be reliably analyzed due to the excessive noise induced by MRI recordings, especially for the English dataset. Without the output voicing measure, it is difficult to assess how these articulatory dimensions contribute to the voicing/devoicing tasks in different languages. Previous studies suggest that the appearance of closure voicing is quite variable in English, but relatively robust in French [1], [3]. Thus, one may expect that voicing-sustaining mechanisms rather than devoicing mechanisms differ between these two languages. However, we found no evidence for more voicing-sustaining mechanisms in French than English. Language-specific differences seem to lie primarily in the devoicing aspect. It is unclear how these differences are related to the closure voicing differences observed across languages. One possibility is that other unexamined articulatory dimensions, such as glottal states, tongue body height, timing among gestures, and the respiratory gestures that modulate subglottal/transglottal pressure, play a role.

## 5. REFERENCES


[1] P. A. Keating, "Phonetic and phonological representation of stop consonant voicing," *Language (Baltim).*, vol. 60, no. 2, p. 286, 1984, doi: 10.2307/413642.

[2] T. Cho, D. H. Whalen, and G. Docherty, "Voice onset time and beyond: Exploring laryngeal contrast in 19 languages," *J. Phon.*, vol. 72, pp. 52–65, Jan. 2019, doi: 10.1016/j.wocn.2018.11.002.

[3] L. Davidson, "Variability in the implementation of voicing in American English obstruents," *J. Phon.*, vol. 54, pp. 35–50, Jan. 2016, doi: 10.1016/j.wocn.2015.09.003.

[4] L. Goldstein and C. P. Browman, "Representation of voicing contrasts using articulatory gestures," *J. Phon.*, vol. 14, no. 2, pp. 339–342, 1986, doi: 10.1016/s0095-4470(19)30662-x.

[5] C. P. Browman and L. M. Goldstein, "Towards an articulatory phonology," *Phonol. Yearb.*, vol. 3, pp. 219–252, 1986, doi: 10.1017/s0952675700000658.

[6] A. Löfqvist, "Laryngeal mechanisms and interarticulator timing in voiceless consonant production," in *Producing Speech: Contemporary Issues for Katherine Safford Harris*, 1995, pp. 99–116.

[7] R. P. Dixit, "Glottal gestures in Hindi plosives," *J. Phon.*, vol. 17, no. 3, pp. 213–237, Jul. 1989, doi: 10.1016/s0095-4470(19)30431-0.

[8] J. R. Westbury, "Enlargement of the supraglottal cavity and its relation to stop consonant voicing," *J. Acoust. Soc. Am.*, vol. 73, no. 4, pp. 1322–1336, 1983, doi: 10.1121/1.389236.

[9] M. Rothenberg, "The breath-stream dynamics of simple-released-plosive production," 1966. doi: 10.1016/0024-3841(71)90101-x.

[10] M. Halle and K. N. Stevens, "A note on laryngeal features," in *MIT Research Laboratory of Electronics Quaterly Progress Report*, 1971, pp. 198–213. doi: 10.1515/9783110871258.45.

[11] F. Bell−Berti, "Control of pharyngeal cavity size for English voiced and voiceless stops," *J. Acoust. Soc. Am.*, vol. 57, no. 2, pp. 456–461, 1975, doi: 10.1121/1.380468.

[12] M. J. Solé, "Articulatory adjustments in initial voiced stops in Spanish, French and English," *J. Phon.*, vol. 66, pp. 217–241, Jan. 2018, doi: 10.1016/j.wocn.2017.10.002.

[13] B. Parrell, "Dynamical account of how /b, d, g/ differ from /p, t, k/ in Spanish: Evidence from labials," *Lab. Phonol.*, vol. 2, no. 2, Oct. 2011, doi: 10.1515/labphon.2011.016.

[14] R. Kagaya, "A fiberscopic and acoustic study of the Korean stops, affricates and fricatives," *J. Phon.*, vol. 2, no. 2, pp. 161–180, Mar. 1974, doi: 10.1016/s0095-4470(19)31191-x.

[15] A. Löqvist, T. Baer, N. S. McGarr, and R. S. Story, "The cricothyroid muscle in voicing control," *J. Acoust. Soc. Am.*, vol. 85, no. 3, pp. 1314–1321, 1989, doi: 10.1121/1.397462.

[16] R. S. McGowan and E. L. Saltzman, "Incorporating aerodynamic and laryngeal components into task dynamics," *J. Phon.*, vol. 23, no. 1–2, pp. 255–269, 1995, doi: 10.1016/S0095-4470(95)80047-6.

[17] I. G. Mattingly, "The global character of phonetic gestures," *J. Phon.*, vol. 18, no. 3, pp. 445–452, 1990, doi: 10.1016/s0095-4470(19)30372-9.

[18] K. Isaieva, Y. Laprie, J. Leclère, I. K. Douros, J. Felblinger, and P. A. Vuissoz, "Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers," *Sci. Data*, vol. 8, no. 1, pp. 1–9, Oct. 2021, doi: 10.1038/s41597-021-01041-3.

[19] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, 2012, doi: 10.1109/TASL.2011.2172425.

[20] S. Narayanan *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1307–1311, 2014, doi: 10.1121/1.4890284.

[21] A. Katsamanis, M. P. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign : Robust long speech-text alignment," *Proc. Work. New Tools Methods Very Large Scale Res. Phonetic Sci.*, pp. 28–31, 2011.

[22] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Trans. Med. Imaging*, vol. 28, no. 3, pp. 323–338, Mar. 2009, doi: 10.1109/TMI.2008.928920.

[23] M. Oh and Y. Lee, "ACT: An Automatic Centroid Tracking tool for analyzing vocal tract actions in real-time magnetic resonance imaging speech production data," *J. Acoust. Soc. Am.*, vol. 144, no. 4, pp. EL290–EL296, 2018, doi: 10.1121/1.5057367.

[24] M. Tiede, "MVIEW: Multi-channel visualization application for displaying dynamic sensor movements." 2010.

[25] D. Bates, M. Maechler, B. Bolker, and S. Walker, "The lme4 Package. R package version 1.1-25." 2020.

[26] J. R. Westbury and P. A. Keating, "On the naturalness of stop consonant voicing," *J. Linguist.*, vol. 22, no. 1, pp. 145–166, 1986, doi: 10.1017/S0022226700010598.

[27] S. Ahn, "The role of tongue position in laryngeal contrasts: An ultrasound study of English and Brazilian Portuguese," *J. Phon.*, vol. 71, pp. 451–467, Nov. 2018, doi: 10.1016/j.wocn.2018.10.003.