Original Article

Voice and Emphasis in Arabic Coronal Stops: Evidence for Phonological Compensation

Vladimir Kulikov

Qatar University, Qatar

Abstract

The current study investigates multiple acoustic cues—voice onset time (VOT), spectral center of gravity (SCG) of burst, pitch (F0), and frequencies of the first (F1) and second (F2) formants at vowel onset—associated with phonological contrasts of voicing and emphasis in production of Arabic coronal stops. The analysis of the acoustic data collected from eight native speakers of the Qatari dialect showed that the three stops form three distinct modes on the VOT scale: [d] is (pre)voiced, voiceless [t] is aspirated, and emphatic [tˁ] is voiceless unaspirated. The contrast is also maintained in spectral cues. Each cue influences production of coronal stops while their relevance to phonological contrasts varies. VOT was most relevant for voicing, but F2 was mostly associated with emphasis. The perception experiment revealed that listeners were able to categorize ambiguous tokens correctly and compensate for phonological contrasts. The listeners' results were used to evaluate three categorization models to predict the intended category of a coronal stop: a model with unweighted and unadjusted cues, a model with weighted cues compensating for phonetic context, and a model with weighted cues compensating for the voicing and emphasis contrasts. The findings suggest that the model with phonological compensation performed most similar to human listeners both in terms of accuracy rate and error pattern.

Keywords

VOT, voicing, emphasis, cue, Arabic, Qatari dialect

Introduction

Mapping acoustic cues to phonological categories is an important and complicated area in phonology and speech perception because links between cues and features are usually multidimensional. Features are typically encoded by several cues, and cues are often used to encode more than one feature (Repp, 1983; Lisker, 1986; Nearey, 1989). In the process of categorization, listeners usually

Corresponding author:
Vladimir Kulikov, Department of English Literature & Linguistics (DELL), Qatar University, P.O. box 2713, Doha, 2713, Qatar.
Email: vkulikov@qu.edu.qa

have to deal with two issues. First, they are to *parse* an acoustic signal (Fowler, 1984; Gow, 2003) and attribute *combined* acoustic cues to sources, such as gestures (Fowler, 1984; Fowler & Brown, 2000) or phonetic features (Gow, 2003; Cole et al., 2010). Since each cue provides an estimate of the relevant part of the acoustic signal, listeners assign some weight or importance to each cue in order to get an accurate estimate of the whole category. Although researchers largely agree that the weight of a cue is a function of its reliability to predict a category (e.g., Nearey, 1990; Toscano & McMurray, 2010, among others), particular mechanisms of cue weighting in categorization are a matter of debate. Second, listeners are to deal with ambiguity and *compensate* for coarticulation or absence of cues (Mann, 1980). Research on compensation (Fowler, 2006; Gaskell & Marslen-Wilson, 1996; Lotto et al., 1997; Lotto & Kluender, 1998; Viswanathan et al., 2010) suggests that listeners are likely to recover the intended category by assigning ambiguity in an acoustic signal to coarticulation. However, it is not clear how compensation affects cue weighting. Ambiguous cues may be evaluated as less reliable and assigned smaller weight. But when a cue is missing in a signal, listeners may compensate for it by assigning greater weight to other cues.

Most studies of speech perception investigated the mechanisms of cue integration, cue parsing, and compensation using cases in which the same cue is shared by two adjacent segments, or CV/VC/CC diphones: F3 as a cue to place of articulation in *rd/rg/ld/lg* clusters (Mann, 1980), nasal airflow in *VN* diphones (Gow, 2003; Fowler & Brown, 2000), vowel duration as a primary cue to voicing of a stop and a secondary cue to vowel height in *bat/bad/bet/bed* syllables (Nearey, 1990), F1 as a primary cue to vowel height and a secondary cue to stop voicing in *hVt/d* syllables (Nearey, 1997), lip rounding in si/sy/ʃi/ʃy syllables (Nearey, 1992; Smits, 2001), or VOT as a primary cue to stop voicing and secondary to vowel duration in *p/bV* syllables (Miller et al., 1986).

In each of these cases, the same cue contributes to categorization of both segments in a diphone, which includes some sort of compensation mechanism or *trading relation* between cues (Repp, 1983). The task becomes more challenging when listeners have to deal with parsing of multiple cues to multiple categories. Research on vowel-to-vowel coarticulation (Cole et al., 2010; McMurray et al., 2011) and on categorization of English fricatives (McMurray & Jongman, 2011) suggests that ambiguity in cues can (at least partially) be resolved by removing contextual information about talkers and neighboring segments. There is little research on cases where multiple cues are used to categorize single segments that belong to multiple phonological contrasts. An example of such a case is Arabic coronal stops د [d], ت [t], and ط [t], which are crosscut by two phonological dimensions: voicing and emphasis. Both contrasts are linked to essentially the same cues: VOT, mean frequency of burst (SCG of burst), fundamental frequency (F0), and frequencies of the first (F1) and second (F2) formants of the following vowel. One might suggest that these cues would differ in terms of their relevance to the two phonological dimensions. For example, VOT would be important to distinguish voicing in stops (Lisker & Abramson, 1964; Yeni-Komshian et al., 1977), but F2 or F1 would be more relevant for the contrast in emphasis (Jongman et al., 2011). However, if one of these cues is missing or becomes ambiguous, proper categorization of stops might be a challenging task. It is not clear how acoustic cues are parsed and what kind of compensation mechanism might be used in this case.

One of the possibilities not often discussed in the literature is that listeners might use some kind of "phonological," contrast-specific compensation. Unlike compensation for coarticulation, in which listeners prefer to assign ambiguity in acoustic signal of a segment to articulation of the neighboring segment, "phonological" compensation is listeners' preference to resolve ambiguity in the acoustic signal in favor of a neighboring category in the same phonological dimension. Previous research showed that listeners can use lexical and phonological knowledge to compensate for ambiguity resulting from phonetic or phonological assimilation (Gaskell & Marslen-Wilson, 1996;

McMurray et al., 2009). This knowledge was shown to be language specific and learned from experience (Darcy et al., 2009).

The computational mechanism of such compensation can be analogous to cue weighting utilized in some models of speech perception (Nearey, 1990; Toscano & McMurray, 2010). If the cue is linked to more than one contrast, listeners may weigh the same cue for each contrast in a segment differently and use this difference to compensate for absence of other cues or context. In order to be retrieved and used in perception, the difference must be learned from previous experience. Toscano and McMurray (2010) argued that in order to be learned, the development of speech categories "must be at least partially category independent" (p. 438). Therefore, contrast-specific cue weighting can be viewed as part of phonological compensation, in which listeners may use expectations derived from *phonological knowledge* in addition to expectations derived from *phonetic context* (Smits, 2001).

To investigate mechanisms of "phonological" compensation, the current paper looks into acoustic properties and categorization of three coronal stops [t], [d], and [ṭ] specified for phonological contrasts of voicing and emphasis in a vernacular Arabic dialect of Qatar. Previous studies of voicing and emphasis in the world's languages suggest that the two contrasts utilize the same or similar acoustic cues, for example, voice onset time, or VOT (Lisker & Abramson, 1964), release burst spectrum (van Alphen & Smits, 2004; Jongman et al., 2011), fundamental frequency (F0) (Ohde, 1984; Kingston & Diel, 1994), F1 frequency (Westbury, 1983; Kingston & Diel, 1994; Jongman et al., 2011) and F2 frequency on vowel onset (Jongman et al., 2011; Zawaydeh & de Jong, 2002). This dialect provides a convenient situation to evaluate contrast-specific weighting of cues and contrast-specific compensation in categorization. In particular, the paper focuses on the case when the three categories of Arabic coronal stops are produced with potentially ambiguous short-lag VOT values. To the best of our knowledge, little or no research has been done in this area.

2   Voicing and emphasis in Qatari Arabic

Multiple cues have been reported to encode the phonological contrasts of voicing and emphasis in Arabic coronal stops (Jongman et al., 2011; Khattab et al., 2006). Similar to other Khaleeji (Gulf) dialects, Qatari Arabic has three contrastive stops at coronal place of articulation: voiced [d], voiceless [t], and voiceless emphatic [ṭ] (Feghali, 2008). Defined broadly as "voiced" or "voiceless" in phonological descriptions of Arabic (e.g., Watson, 2002), the three categories of stops form three distinct modes on the VOT scale. Voiced and voiceless coronal stops also differ in VOT. Voiced [d] is produced with voice lead ($M = -58$ ms), voiceless [t] is aspirated, with long-lag VOT ($M = 54$ ms) (Kulikov, 2020), and emphatic [ṭ] has short-lag VOT ($M = 16$ ms) (Kulikov et al., 2020). But unlike in languages with a three-way voicing contrast, for example Thai or Eastern Armenian (Lisker & Abramson, 1964), the three modes on the VOT scale in Qatari Arabic distinguish two phonological contrasts: voicing and emphasis. The voiceless unaspirated stop [ṭ] is not only a voiceless category but is also an emphatic category.

Emphasis in Arabic coronal obstruents is a phonological contrast in secondary place of articulation (Ladefoged & Maddieson, 1996). The primary cues to emphasis are spectral. Emphatic coronal obstruents are articulated with a retracted tongue root and/or raised tongue back (McCarthy, 1994), which causes lowering spectral mean of a consonant (Jongman et al., 2011), including lower SCG of burst of a stop (Kulikov et al., 2020). Emphatic articulation is typically spread onto an adjacent vowel (Zawaydeh & de Jong, 2002) changing its spectral characteristics. Acoustic correlates of emphasis include raising of F1 and lowering of F2 frequencies of the vowel, which usually prevail during the entire duration of the vowel. Importantly, emphasis in a consonant is perceived predominantly through the emphatic quality of the vowel because the differences between the

spectra of plain and emphatic consonants are quite small (Jongman et al., 2011). Similar to other vernacular dialects of Arabic, cues to emphatic consonants in Qatari Arabic include lower SCG of stop bust, as well as raised F1 and lowered F2 on the following vowel (Kulikov et al., 2020). Crucially, these cues also encode the contrast in voicing. Kulikov (2020) reports that vowels following voiced [d] have lower F1 than vowels following voiceless [t] (p. 171). SCG of burst is also lower in voiced stops (Kulikov, 2016).

The two types of cues for emphasis—temporal and spectral cues—are not independent of one another. There seems to be a trading relation between VOT and spectral cues in Arabic emphatic stops. In on-going sound changes among female speakers that result in loss of emphasis, the lower degree of emphasis on a vowel (measured as lower F2 value) has been shown to correlate with longer VOT in a stop (Khattab et al., 2006).

Being phonologically distinct, the three categories of coronal stops reveal an overlap on the VOT scale. In addition to unaspirated [t̪] produced with short-lag VOT, phonologically voiced [d] can be produced without prevoicing, that is, they can have short-lag VOT, and some tokens of [t] have shorter duration of aspiration (Kulikov, 2020). Disambiguation of these stops might require use of secondary cues to voicing—SCG of burst, F0, F1, or F2. It is noteworthy that these cues are also important for perception and production of emphasis in Arabic. Therefore, parsing cues to voicing and emphasis may include a compensation mechanism to adjust cue weights for the two contrasts. For example, shorter VOT may not only disfavor a voiceless stop but also favor emphasis.

The following experiments provide evidence for acoustic properties of the coronal stops in Qatari Arabic and listeners' preferences in segment-to-segment compensation in the absence of rich context. In Experiment I, I looked into the distribution of the three stop categories on the VOT scale and investigated acoustic cues that distinguish voiced, voiceless, and emphatic coronal stops. Experiment II presents perception patterns in human listeners' identification of stops that fall into the ambiguous VOT interval between 0 and 40 ms. As all these tokens had short-lag VOT and were produced before the same vowel, I was looking for and found a bias, or preference in disambiguation, toward a particular stop category that cannot be explained by coarticulation. I argue that the bias can be explained as a type of compensation for "phonological context." Finally, listeners' perception was tested against three models of cue integration that attempted to account for "phonological" compensation.

3 Experiment I: Coronal stops in Qatari Arabic

3.1 Method

The production data were collected from eight native speakers (females) of Qatari Arabic, who were undergraduate students at Qatar University. They were born to families who belong to the original Qatari tribes. None of them reported speech or hearing disorders. The participants were asked to speak in their native dialect as if they were talking to family and friends. The instructions were delivered in colloquial Arabic by Arabic-speaking research assistants.

The recordings were made in a quiet room using a Shure WH30XLR microphone and a portable Marantz PND661 MKII recorder. The participants pronounced (read) words (n = 25) with word-initial voiced *d* (n = 9), voiceless *t* (n = 9), and emphatic *t̪* (n = 7) stops presented to them in Arabic orthography. The list of words used in the experiment is given in the Appendix. Each target word was pronounced six times in a carrier phrase *Qaalet . . . marratain* "She said . . . one more time."

The recorded materials were evaluated by two native speakers of Qatari Arabic for naturalness. Four tokens were discarded due to mispronunciation. The total of 1196 tokens were prepared for
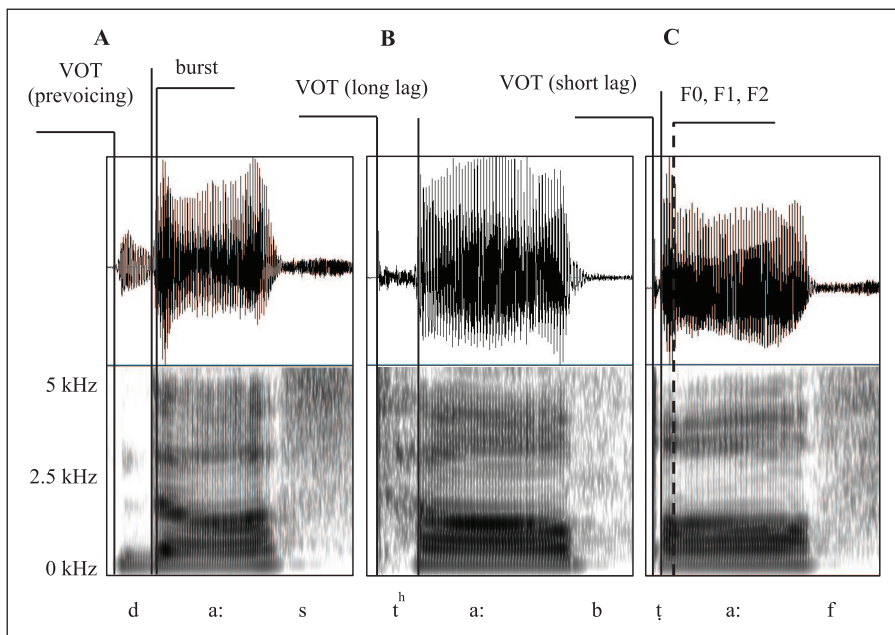
Figure 1. The waveforms and spectrograms of a voiced (A), a voiceless (B), and a voiceless emphatic (C) coronal stop in Qatari Arabic: prevoiced [d] in daas "stepped," voiceless aspirated [tʰ] in taab "repented," voiceless unaspirated emphatic [ṭ] in ṭaaf "ignore."

acoustic analysis. The segment boundaries were set manually in PRAAT (Boersma & Veenink, 2015). VOT was measured as timing between the stop release and the onset of voicing. Both waveforms and spectrograms were used to identify the beginning of glottal pulses. Spectral cues, such as fundamental frequency (F0) and frequency of the first (F1) and second (F2) formant at vowel onset, as well as mean SCG of burst, were measured to evaluate the glottal state and degree of emphasis (pharyngealization) of stop articulation during the release. The landmarks for measurements are summarized in Figure 1. For subsequent analysis of stops in the ambiguous VOT range between 0 and 40 ms, 100 voiced [d]s, 120 voiceless [t]s and 120 emphatic [ṭ]s were randomly selected out of 565 tokens that overlapped on the VOT scale.

3.2 Results I: Contrastive categories of coronal stops

Observation of distributions of VOT values (Figure 2A) revealed that voiceless, voiced, and emphatic stops in Qatari Arabic form three distinct modes. In Lisker and Abramson's (1964) terms, they would present three laryngeal categories: (pre)voiced [d], voiceless aspirated [t], and voiceless unaspirated [ṭ]. The voiceless unaspirated category, however, does not form a separate "voicing" category in Arabic phonology. It marks an emphatic category [ṭ], which, in addition, would be typically identified by formant frequencies of the neighboring vowel, F2 being the most characteristic cue to emphasis (Zawaydeh & de Jong, 2002). Observation of distributions of these values (Figure 2B) suggested that there is a boundary at approximately 1500 Hz, with lower values marking the emphatic category. Both cues revealed some overlap between contrastive categories suggesting possible trade-in relation between cues in production and compensation in perception of the stop categories.
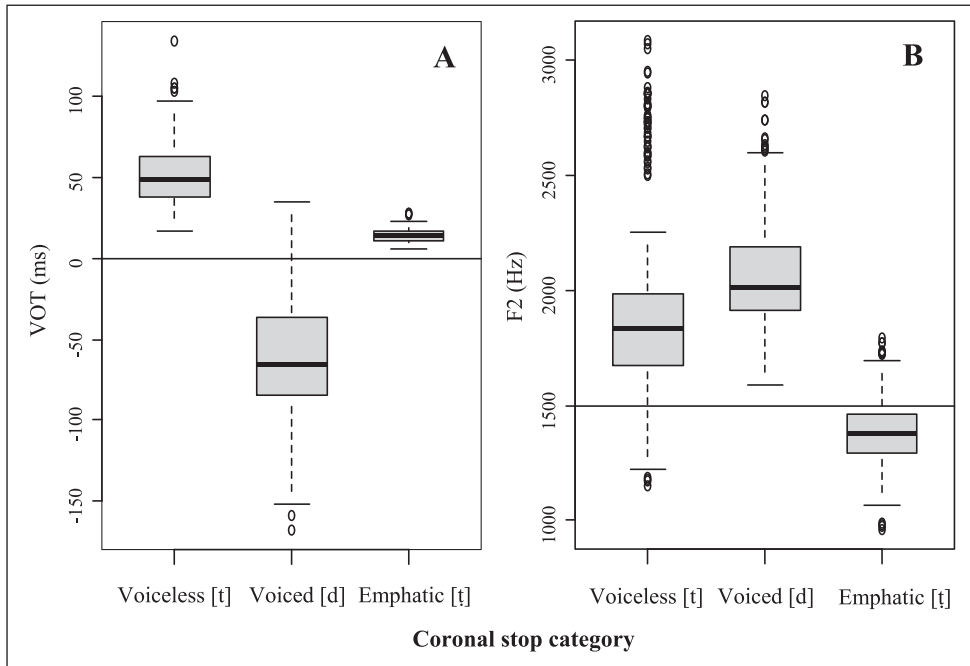
Figure 2. Boxplots of (A) VOT and (B) F2 values of coronal stops [d], [t], and [ṭ] in Qatari Arabic.

3.3 Results II: Cues to voicing and emphasis in coronal stops

Analysis of the cues included several stages. First, the five cues to voicing and emphasis—VOT, SCG of burst, F0, F1, and F2—were tested for the effect of stop category (voiceless [t], voiced [d], and voiceless emphatic [ṭ]). The effect of stop identity was evaluated using linear mixed-effects models in the lme4 package (Bates et al., 2015) in R (R Core Team, 2019). Each of the five acoustical cues was a dependent variable. Stop category was used as a fixed factor; talker and item were used as random factors. As independent variables were manipulated within-subjects, it was plausible to use both random slope and random intercept for the talker and item (Barr et al., 2013) to adjust for talker and item variation in production of the acoustic cues for different categories of stops in different types of syllables. Although Likelihood ratio tests showed that the maximal model achieved a better fit than the models without random slope for talker, $\chi^2(5) = 172.53$, $p < .0001$, and item, $\chi^2(5) = 12.54$, $p < .05$, the improvement was found only for VOT. Adding random slopes to the models evaluating spectral cues did not improve model fit. Therefore, random slopes were excluded from the final model for the benefit of model convergence (Matuschek et al., 2017) and consistency of representation of the results for each cue.

3.3.1 Effect of stop category on each cue. Table 1 summarizes the results of the fixed effect of stop category for the five cues, and the effects are plotted in Figure 3. The *p*-values for factor levels were calculated using the lmerTest package (Kuznetsova et al., 2017). The stop categories were coded using simple coding so that the reference category (intercept) was voiceless [t], and the parameter estimates for voiced [d] and emphatic [ṭ] (lines 2 and 3 for each cue in Table 1) were evaluated against it. As the model does not provide a procedure to compare the latter stops to each other, the comparison between voiced [d] and emphatic [ṭ] (line 4 for each cue in Table 1) was done using a pairwise comparison of EMMs in the emmeans package (Lenth, 2020).

Table 1. Summary of mixed-effects linear models examining an effect of stop category (voiceless [t], voiced [d], emphatic [ṭ]) on each cue. The reference category is voiceless [t].

Cue	Level	Estimate	Std. Error	T value	Pr (> |t|)
VOT	1. Intercept	51.2	4.2	12.29	< .0001
	2. [d]	−104.8	3.5	−30.11	< .0001
	3. [ṭ]	−36.9	3.7	−9.94	< .0001
	4. [d] vs. [ṭ]	−67.9	3.7	−18.23	< .0001
	Residual	750.3			
	Item (intercept)	38.8			
	Talker (intercept)	90.3			
SCG of burst	1. Intercept	1987.1	202.5	9.81	< .0001
	2. [d]	−1190.4	159.9	−7.44	< .0001
	3. [ṭ]	−864.5	170.9	−5.06	< .0001
	4. [d] vs. [ṭ]	−325.9	116.4	−2.81	.021
	Residual	597121			
	Item (intercept)	102533			
	Talker (intercept)	225911			
F0	1. Intercept	235.1	6.7	35.28	< .001
	2. [d]	−15.5	1.9	−7.87	< .001
	3. [ṭ]	−12.9	2.1	−6.12	< .001
	4. [d] vs. [ṭ]	−2.6	2.1	−1.24	.441
	Residual	170.8			
	Item (intercept)	13.9			
	Talker (intercept)	339.9			
F1	1. Intercept	575.5	37.4	15.38	< .001
	2. [d]	−96.8	48.9	−1.98	.060
	3. [ṭ]	83.2	52.3	1.59	.125
	4. [d] vs. [ṭ]	−180.0	52.3	−3.44	.006
	Residual	4027			
	Item (intercept)	10699			
	Talker (intercept)	1599			
F2	1. Intercept	1888.3	84.4	22.35	< .001
	2. [d]	191.5	117.3	1.64	.136
	3. [ṭ]	−505.9	123.2	−4.10	< .001
	4. [d] vs. [ṭ]	697.4	125.2	5.57	< .0001
	Residual	16966			
	Item (intercept)	61387			
	Talker (intercept)	4138			

The effect of stop category was significant for most cues. VOT was the longest in voiceless [t], ($M = 51$ ms); it averaged 14 ms in emphatic [ṭ], and it was negative averaging—54 ms in voiced [d]. SCG of burst was the highest in voiceless [t], ($M = 1987$ Hz); it decreased in emphatic [ṭ] averaging 1122 Hz and in voiced [d] averaging 797 Hz. F0 at vowel onset was the highest after voiceless [t] ($M = 235$ Hz); it decreased after emphatic [ṭ] ($M = 220$ Hz) and after voiced [d] (M = 219 Hz). The latter were not significantly different from one another ($p = .441$). F1 was on average the highest after emphatic [ṭ] ($M = 659$ Hz); it was lower after voiceless [t] ($M = 576$ Hz) but the difference was not significant, and the lowest after voiced [d] ($M = 479$ Hz). F2 was after voiced [d] ($M = 2267$ Hz); it was slightly lower after voiceless [t] ($M = 1888$ Hz) but the difference was not significant, and the lowest F2 was found after emphatic [ṭ] ($M = 1383$ Hz).
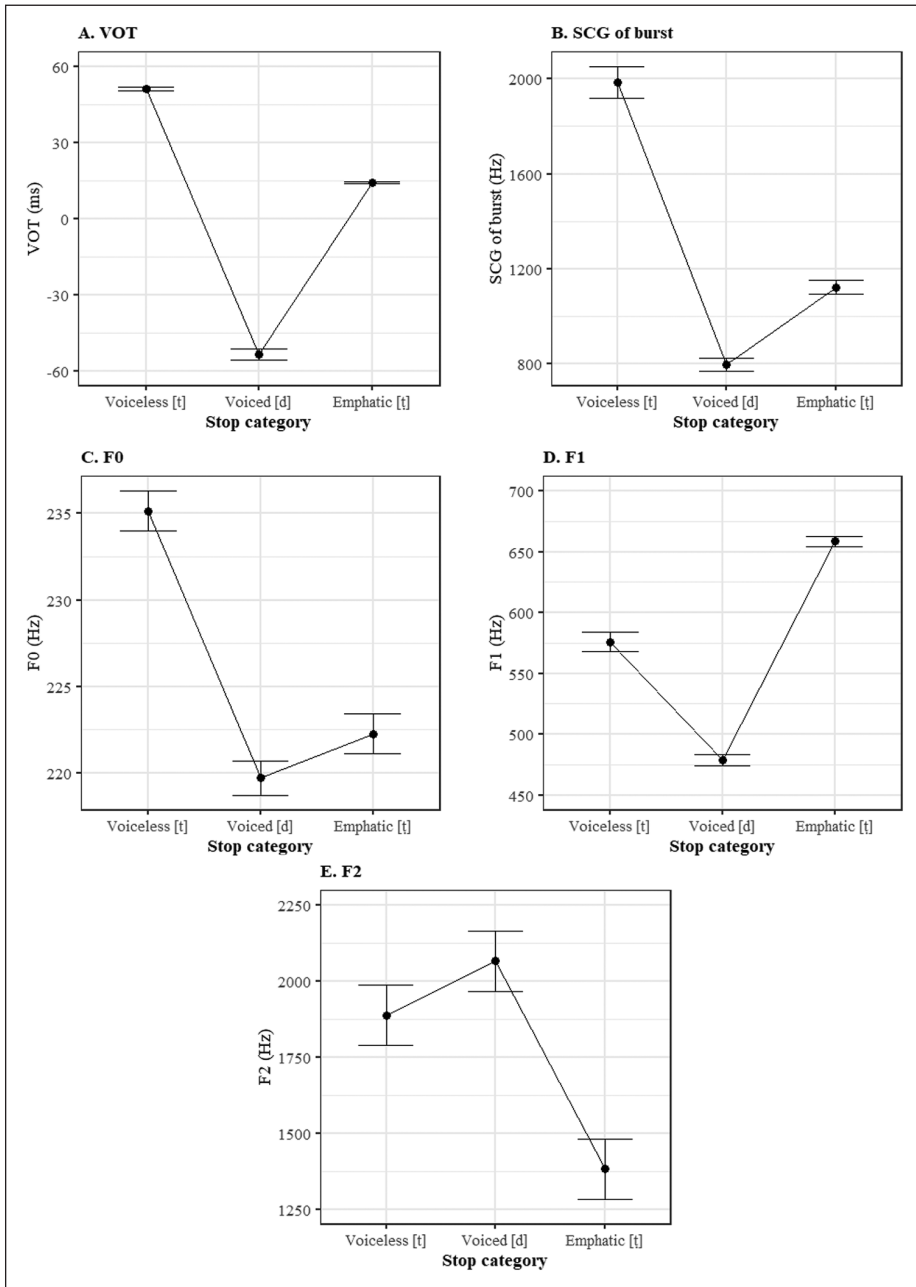
Figure 3. Effect plots for five acoustic cues (VOT, SCG of burst, F0, F1, and F2) to coronal stops [t], [d] and [ṭ] in Qatari Arabic.

3.3.2 Variability in cues. The previous analysis revealed significant differences in production of cues for each stop category; however, it did not show how important each cue was to distinguish a stop category. Nor did it account for the fact that the three categories were crosscut by two phonological dimensions: voicing (voiced [d] vs. voiceless [t] and [ṭ]) and emphasis (emphatic [ṭ] vs.

Table 2. Summary of regression analyses examining effects of talker, vowel, and the phonological contrasts on each cue. R^2_{change} values are shown.

Cue	Talker	Vowel	Stop identity	
			Voicing	Emphasis
	$(df = 7, 1188)$	$(df = 4, 1184)$		
			$(df = 1, 1183)$	$(df = 1, 1182)$
VOT	.029***	.099***	.564***	.043***
SCG of burst	.175***	.106***	.113***	.039***
F0	.564***	.033***	.043***	.024***
F1	.072***	.514***	.138***	.001^
F2	.023***	.474***	.119***	.141***

Note: significance levels: $^\wedge - p = .05$, $*** - p < .001$.

non-emphatic [t] and [d]). Evaluation of contextual variability in each cue is a crucial condition for categorization and compensation. If a cue is invariant, listeners may not need to weight or adjust it. Categorization under this scenario may not require compensation of any kind. Possible invariance of cues and their relevance to categorization of voicing and emphasis was evaluated in a series of hierarchical linear regression models. Each cue was a dependent variable; the independent variables were dummy codes for talker identity, vocalic context, and the stop identity split between two phonological contrasts. R^2_{change} was used as a measure of effect size (Cohen & Cohen, 1983), revealing the proportion of variance explained by a predictor variable in question at each step. The effect was considered small if R^2_{change} was smaller than .05, medium if it was between .05 and .15, and large if it was greater than .15 (see Table 2).

I adopted the invariance criterion used in McMurray and Jongman (2011). The cue is invariant if contextual effects of talker and vowel are small or medium, and the effect of a phonological feature is large. Hierarchical regression provides a computational mechanism to remove parts of variance and evaluate each source of variance separately. First, the effects of talker and context were partialed out by adding seven dummy codes that represented variance in eight participants, and four dummy codes that represented vocalic context: one for length of a following vowel (short or long), one for number of syllables (one or two), and two for vowel category (a, i, u) in target words. The effect sizes were consistent with the relative amount of variance for the random effects of talker and item observed in the mixed-effects models (section 3.3.1). The talker accounted for a significant amount of the variance in all cues, but the size varied from being small for VOT and F2 to medium for F1, and small for SCG of burst and F2. Vocalic context also accounted for a significant amount of the variance in all cues. The effect size was small for F0, medium for VOT and SCG of burst, and large for F1 and F2.

Next, the dummy codes for voicing and emphasis were added to the regression model separately. Codes for voicing were added first to account for the fact that prevoicing in most voiced stops must signal the presence of voicing at a relatively early stage. For the voicing contrast, the largest effect size was obtained for VOT, the effect sizes were medium for SCG of burst, F1, and F2, and the smallest effect size was obtained for F0. For the emphasis contrast, the largest effect size was obtained for F2, and small effect sizes were found for SCG of burst, F0, and F1.

Each phonological contrast revealed one primary acoustic correlate: VOT was the most important cue to voicing, and F2 was the primary cue to emphasis. The effect of phonological contrast on each of them was larger than the sum of the effects for other cues. But essentially only VOT met the definition of an invariant cue for the voicing contrast: it had a medium combined effect of context ($R^2 = .128$) and a very large effect of the phonological feature ($R^2 = .564$). F2 approached the
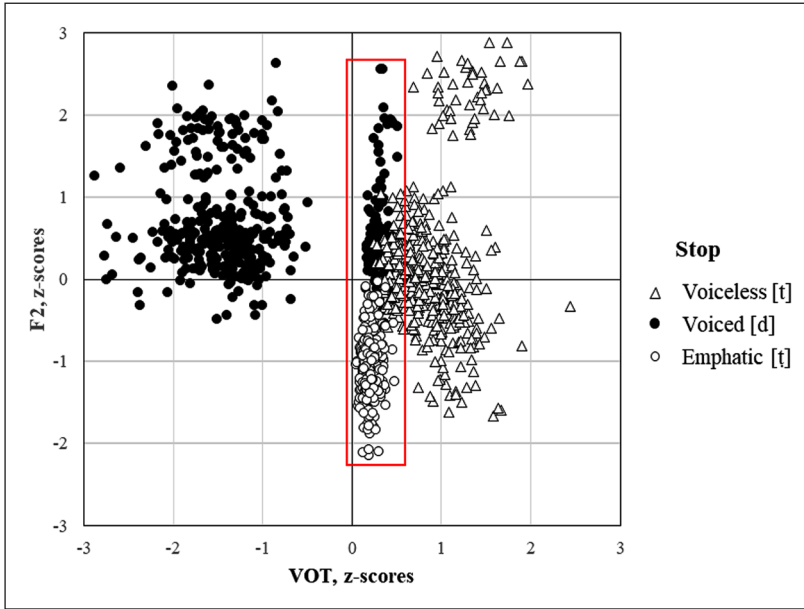
Figure 4. Distributions of VOT and F2 values in voiceless [t], voiced [d], and emphatic [t̪] in Qatari Arabic. The red square represents the ambiguous area between 0 and 40 ms on the VOT scale.

definition of an invariant cue to the emphasis contrast: it had the relatively large effect of a phonological feature ($R^2 = .141$), but the combined effect of context was also very large ($R^2 = .497$).

3.3.3 Interim summary. The acoustic analyses of the cues to voicing and emphasis revealed significant differences between the categories of coronal stops in each of the five cues. The coronal stops formed distinct categories on the VOT scale: [d] was mostly (pre)voiced, [t̪] was voiceless unaspirated, and [t] was voiceless aspirated. In addition, the voicing and emphasis contrasts in coronal stops were maintained in spectral cues. The contrast between voiceless [t] and voiced [d] was manifested as difference in SCG of burst and F0. Emphatic [t̪] was different from non-emphatic stops in F2 and SCG of burst, and it was different from voiced [d] in F1.

The relevance of cues for each phonological contrast varied. VOT, quite expectedly, was primarily associated with the voicing contrast. F2, on the contrary, was primarily associated with the emphatic category. The other spectral cues—SCG of burst, F0, and F1—had smaller effects of voicing and emphasis. Absence of invariant cues for both contrasts suggests that cue weighting and compensation may be crucial in categorization of Arabic coronal stops, especially in cases when stops are ambiguous. The acoustic analysis of stops overlapping on the VOT scale is presented in the following section.

3.4 Results II: Cues in stops in the ambiguous VOT range

3.4.1 Distribution of VOTs in three categories of stops. Analysis of distributions of VOT in voiced, voiceless, and emphatic alveolar stops in the data revealed a considerable overlap between the three categories within the range 0–40 ms, that is, the range in which voiceless unaspirated stops are typically produced. In addition to voiceless unaspirated emphatic [t̪]s, 23% of Qatari Arabic [d]s were produced without prevoicing. Of voiceless [t]s 25% had shorter duration of aspiration in the range of 35–40 msec. (Figure 4).

The questions to ask at this point are the following: to what extent will overlap on the VOT scale lead to neutralization between the three categories? and to what extent will spectral cues compensate for potential loss of contrast on the VOT scale? Compensation is, in fact, expected because *no* stop category differed from the other two in *all* cues. Voiced [d] and emphatic [tˤ] were different in F1 and F2 frequencies but showed no significant difference in SCG of burst or F0. Voiceless [t] and emphatic [tˤ] were different in SCG of burst, F0, and F2 frequencies but showed no significant difference in F1. Voiced [d] and voiceless [t] were different in SCG of burst, F0, and F1 frequencies; however, they showed no significant difference in F2.

3.4.2 Cues in stops within the ambiguous VOT range. For the analysis of the cues in stops within the ambiguous VOT range, separate mixed-effects linear models for each cue were fit to the data. Stop category was a fixed factor, talker and item were random factors (intercepts). Pairwise comparison of EMMs was used to evaluate the difference between voiced and emphatic stops. Table 3 summarizes the models.

The effect of stop category was significant for all cues. Despite the overlap, the three VOT categories were significantly different from one another, although the means were within the range characteristic of a voiceless unaspirated category. VOT was the longest in voiceless [t] ($M = 34$ ms); it was shorter in emphatic [tˤ] ($M = 14$ ms) and in voiced [d] ($M = 18$ ms). SCG of burst was the highest in voiceless [t] ($M = 2027$ Hz); it decreased in emphatic [tˤ] ($M = 1124$ Hz) and in voiced [d] ($M = 1101$ Hz). The 152 Hz difference between the latter was not significant ($p = .284$). F0 was the highest after voiceless [t] ($M = 235$ Hz); it decreased after emphatic [tˤ] ($M = 224$ Hz) and after voiced [d] ($M = 227$ Hz). The latter were not significantly different from one another ($p = .806$). F1 frequency was on average the highest after emphatic [tˤ] ($M = 659$ Hz); it was lower after voiceless [t], but the difference was not significant, and it was the lowest after voiced [d] ($M = 467$ Hz). F2 frequency was the highest after voiced [d] ($M = 2041$ Hz); it decreased after voiceless [t], and it was the lowest after emphatic [tˤ] ($M = 1383$ Hz).

3.4.3 Interim summary. The results revealed significant differences between three categories in all cues, including VOT. However, mean VOT values for stops in the ambiguous area fell into the range of the voiceless unaspirated category (14–32 ms). Therefore, it is not clear whether these differences have any perceptual effect. Differences in spectral cues were more prominent, and they were similar to the differences reported in Section 4.1 with one exception: F1 was significantly lower ($MD = –125$ Hz) after voiced [d] than after voiceless [t] when the former was produced without prevoicing, suggesting some kind of compensatory mechanism. When VOT is not a reliable predictor of voice, F1 transition may reveal sufficient information about the glottal state of an obstruent (Summerfield & Haggard, 1977). The results suggest that distinction between the categories of overlapping stops is maintained mainly by spectral cues. As three categories of stops are crosscut by two phonological contrasts, each spectral cue may be evaluated in relation to both contrasts. The next section presents a perception study that addresses the question of how cues are weighed by human listeners in a categorization task.

4 Experiment II: Perception of stops in the ambiguous VOT range by human listeners

Accuracy in categorization of coronal stops in the ambiguous VOT range by human listeners was assessed in two conditions: 1) Noise, where only two consonantal cues, VOT and SCG of burst, were directly accessible to listeners, and 2) Vocalic, in which all five cues were available to listeners. In the absence of some cues, listeners are expected to compensate for them. But if phonetic

Table 3. Summary of mixed-effects linear models examining an effect of stop category (voiceless [t], voiced [d], emphatic [ṭ]) on each cue in stops overlapping on a VOT scale. The reference category is voiceless stop [t].

Cue	Level	Estimate	Std. Error	t value	Pr (> |t|)
VOT	1. Intercept	33.5	1.3	25.5	< .001
	2. [d]	−15.1	1.8	−8.3	< .001
	3. [ṭ]	−19.3	1.4	−13.5	< .001
	4. [d] vs. [ṭ]	4.2	1.3	3.2	.024
	Residual	15.7			
	Item (intercept)	.3			
	Talker (intercept)	.9			
SCG of burst	1. Intercept	1339.0	371.6	8.8	< .001
	2. [d]	−599.4	126.7	−4.1	.001
	3. [ṭ]	−432.4	124.9	−4.0	.001
	4. [d] vs. [ṭ]	−152.0	132.2	−1.1	.284
	Residual	413177			
	Item (intercept)	.0			
	Talker (intercept)	887670			
F0	1. Intercept	234.8	7.4	31.7	< .001
	2. [d]	−10.4	3.6	−2.6	.017
	3. [ṭ]	−7.6	3.1	−2.7	.015
	4. [d] vs. [ṭ]	−2.8	3.2	−.3	.806
	Residual	155.6			
	Item (intercept)	8.9			
	Talker (intercept)	625.9			
F1	1. Intercept	699.2	24.5	28.5	< .001
	2. [d]	−177.0	25.5	−6.9	< .001
	3. [ṭ]	−8.7	31.1	−.3	.788
	4. [d] vs. [ṭ]	−185.7	36.3	−5.5	< .001
	Residual	2269			
	Item (intercept)	1133			
	Talker (intercept)	2035			
F2	1. Intercept	1869.0	54.5	34.3	< .001
	2. [d]	139.8	58.3	2.4	.036
	3. [ṭ]	−541.8	72.4	−7.5	< .001
	4. [d] vs. [ṭ]	692.9	81.7	8.5	< .001
	Residual	9157			
	Item (intercept)	6457			
	Talker (intercept)	9343			

context is also impoverished, the compensation mechanism may include addressing to more abstract layers of phonological knowledge about the categories.

4.1 Method

One hundred and thirty-five coronal stops [d], [t], and [ṭ] from Experiment 1 with VOT in the range between 0 and 40 ms were selected for the identification task. They showed reasonable variation in acoustic cues. The files were edited manually to create two types of stimuli: with a noise portion (Noise condition) and with both noise and vocalic portions (Vocalic condition), which yielded a
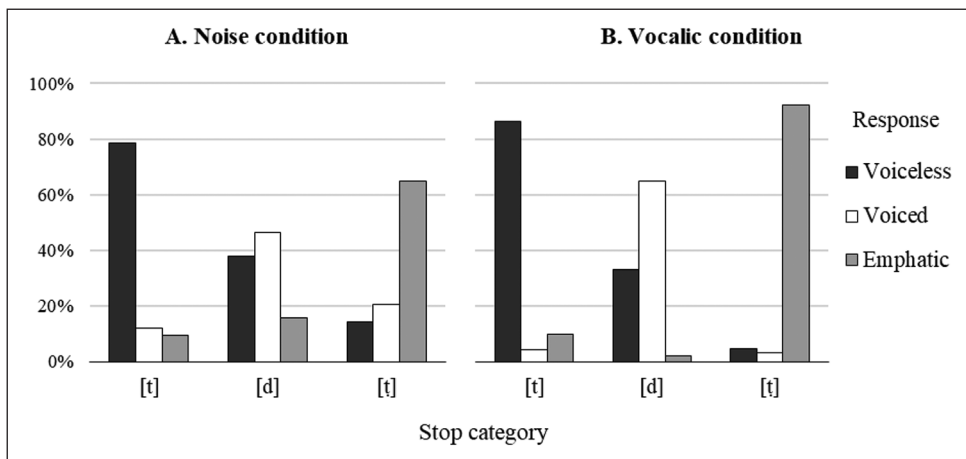
Figure 5. Categorical response patterns of the identification tests for three categories of coronal stops in the noise and vocalic conditions.

total of 270 stimuli. The noise portion included duration of burst from the release point to the beginning of glottal pulsing. The vocalic portion included 20 ms of a vowel after the release capturing the point where the measurements of the vocalic cues (F0, F1, and F2) were performed in the first experiment.

Twenty-eight undergraduates from Qatar University participated in this study, and 14 served in each condition (vocalic vs. noise). All participants belonged to local communities and were native speakers of Qatari Arabic with no known speech or hearing impairment. All of them indicated that they could speak English as a second language. They received a bonus point in an enrolled class for their participation.

The stimuli were played back in random order twice with a 2 sec interval through Direct Sound EX-25 headphones. Listeners responded by circling one of the three options written in standard Arabic orthography (د [d], ت [t], ط [t̪]) on answer sheets. Prior to the actual test, the participants performed a short training test with three non-ambiguous stops to ensure they could hear the difference between the phonemes in Arabic.

4.2 Results I: Categorical responses

The effects of stop identity and condition were evaluated using linear mixed-effects models in the lme4 package (Bates et al., 2015) in R (R Core Team, 2019). Listeners' response score was a dependent variable. Stop category (voiceless, voiced, emphatic) was used as a fixed within-subjects factor; condition (noise, vocalic) was a between-subjects factor; listener and item were random factors; stop category was a random slope for listener to adjust for individual variation in perception of the stimuli of different categories of stops (Barr et al., 2013). The maximal model was selected as it achieved a better fit than the model without a random slope, $\chi^2(18) = 283.69$, $p < .0001$. The model parameters are summarized in Table 4 (Appendix 1). Figure 5 shows the results of the identification test.

In the noise condition, the scores were the highest for voiceless [t] (79%), lower for emphatic [t̪] (65%), and the lowest for voiced [d] (47%). In the vocalic condition, performance significantly improved for voiced [d] and emphatic [t̪], reaching 58% and 85% respectively, but the 7% increase for voiceless [t] was not significant.
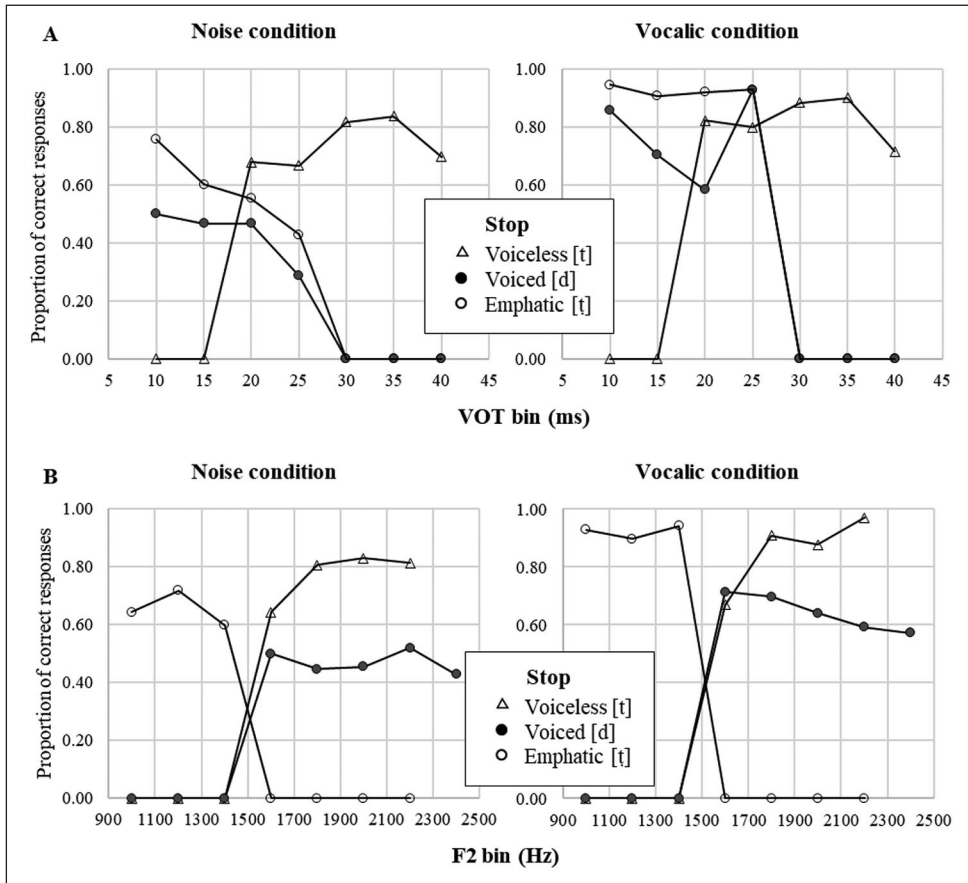
Figure 6. Summary of the identification test for the three coronal stops in Qatari Arabic in the noise and vocalic conditions along primary cues to voicing and emphasis.

The follow-up tests revealed that low performance in voiced [d] was due to confusion with voiceless [t] rather than emphatic [ṭ] in both conditions. Listeners were equally likely to hear [d] as [t] in the noise condition ($Z = -2.1$). In the vocalic condition, this ratio dropped to 50% ($Z = -3.2$) but voiceless [t] was still a more likely competitor than emphatic [ṭ] ($Z = -13.9$). In contrast, listeners were able to categorize the voiceless and emphatic stops without a bias toward a particular category in both noise and vocalic conditions.

4.3 Results II: Compensation

Relationship between listeners' accuracy and acoustic cues was evaluated using linear mixed-effects models in the lme4 package. Listeners' response score was a dependent variable. Stop category (voiceless, voiced, emphatic) and acoustic cue were fixed within-subjects factors; condition (noise, vocalic) was a between-subjects factor, and participant was a random factor. Stop category was used as a random slope for listener to adjust for individual variation in perception of different categories of stops. The model parameters for each cue are summarized in Tables 5–9 (see Appendix 1). Figures 6 and 7 visualize the results.
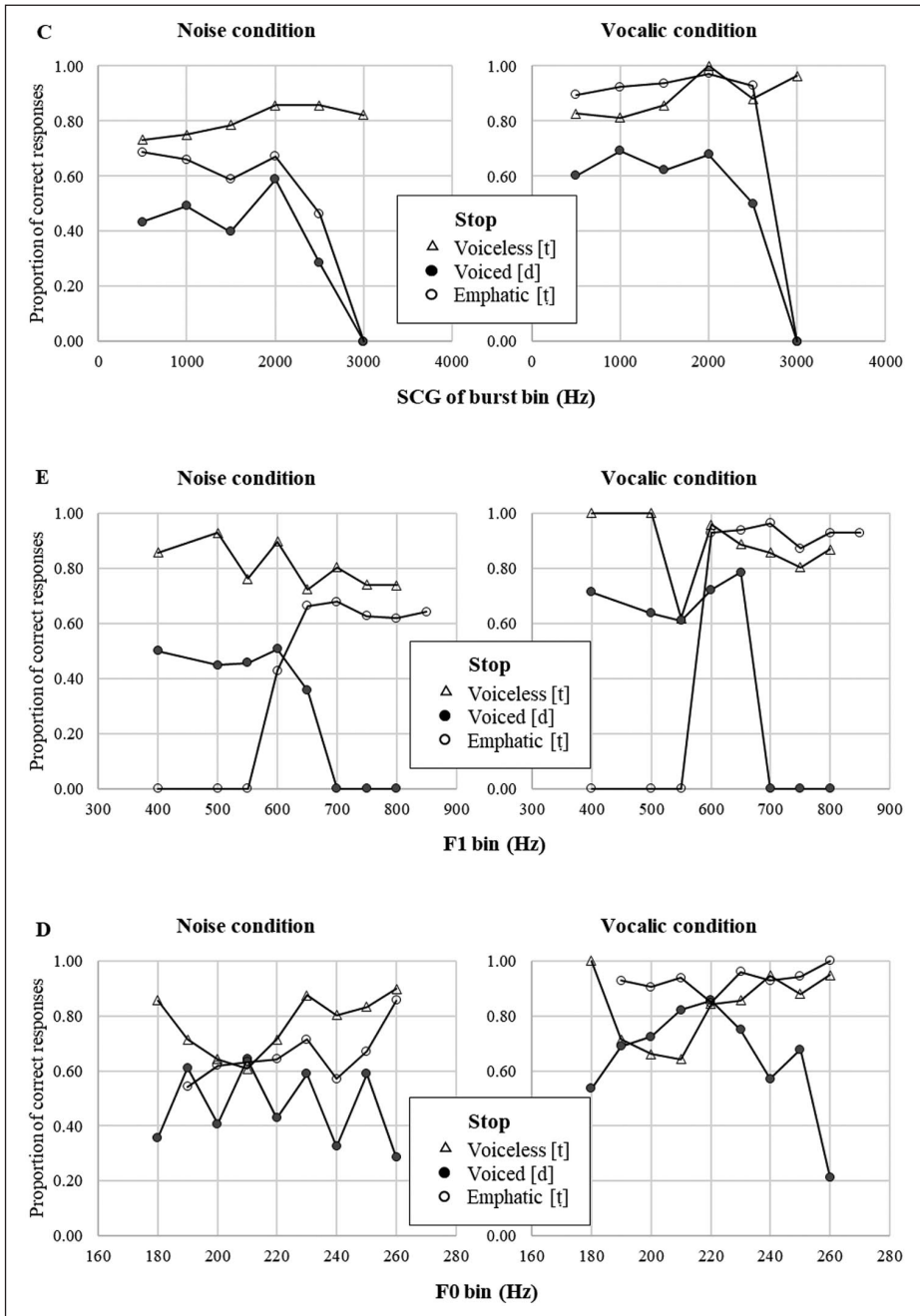
Figure 7. Summary of the identification test for the three coronal stops in Qatari Arabic in the noise and vocalic conditions along secondary cues to voicing and emphasis.

4.3.1 Primary cue to voicing: VOT. For the analysis of VOT, the values were binned in 5 ms intervals (Figure 6A), and mean accuracy scores for each interval were fitted to the linear model (see Table 5). VOT significantly affected accuracy, Log Likelihood: $\chi^2(1) = 20.36$, $p < .05$. Two- and three-way

interactions with condition and stop category indicated that VOT was a more accurate predictor of a stop category in the vocalic condition. Quite predictably, listeners' response to VOT was categorical in distinguishing the voicing contrast between voiceless aspirated [t] and unaspirated [d] and [t]. The categorical boundary was between 20 and 25 ms, and tokens with VOT longer than 30 ms were typically identified as [t] in both conditions. Accuracy in voiced and emphatic stops with ambiguous VOT values decreased (Voiced: $\beta = -.017, p < .05$; Emphatic: $\beta = -.019, p < .01$) as VOT became longer. Pairwise comparison between voiced and emphatic stops revealed that the latter were identified more accurately at 10 ms ($\beta = .256, T = 2.468, p < .05$), 15 ms ($\beta = .184, T = 3.735, p < .05$), and 20 ms ($\beta = .153, T = 2.202, p < .05$). The difference suggests that listeners used some sort of compensation mechanism, resolving ambiguity in unaspirated tokens in favor of emphatic stops.

4.3.2 Primary cue to emphasis: F2.

For the analysis of F2, the values were binned in 200 Hz intervals (Figure 6B), and mean accuracy scores for each interval were submitted to the model (Table 6). F2 significantly affected accuracy, Log Likelihood: $\chi^2(1) = 74.0, p < .0001$. As expected, F2 was a more accurate predictor of a stop category in the vocalic condition ($\beta = .108, p < .05$). Listeners showed a categorical response to F2 to distinguish the contrast between emphatic [t] and non-emphatic [t] and [d]. The categorical boundary was at 1500 Hz; tokens with higher F2 were typically identified as plain [t] or [d]. Interaction with stop category indicated that accuracy in voiced and emphatic stops with ambiguous VOT values decreased (Voiced: $\beta = -.0005, p < .0001$; Emphatic: $\beta = -.0004, p < .05$) as F2 became higher, whereas accuracy of identification of voiceless stops increased ($\beta = .0004, p < .0001$). Pairwise comparison between voiced and voiceless stops revealed that voiceless stops were identified more accurately than voiced stops at 1800 Hz ($\beta = .286, T = 3.848, p < .01$), 2000 Hz ($\beta = .305, T = 4.114, p < .01$), and 2200 Hz ($\beta = .337, T = 4.538, p < .001$). The differences also suggest compensation as listeners resolved ambiguity in plain tokens in favor of voiceless stops.

4.3.3 Secondary cues to voicing and emphasis (SCG of burst, F0, F1).

The frequency values of secondary cues were also binned in 500 Hz intervals for SCG of burst, 10 Hz intervals for F0, and 100 Hz intervals for F1; mean accuracy scores for each interval were submitted to the models (see Tables 7–9). The results show that each cue significantly affected identification accuracy, SCG of burst: Log Likelihood, $\chi^2(1) = 16.62, p < .0001$; F0: Log Likelihood, $\chi^2(1) = 18.64, p < .0001$; F1: Log Likelihood, $\chi^2(1) = 13.14, p < .05$. Vocalic condition improved accuracy for vocalic cues (F0: $\beta = .131, p < .01$; F1: $\beta = .106, p < .05$) but did not affect the consonantal cue (SCG of burst: $\beta = .085, p = .142$). Effects of stop category were significant at all levels for F1 (Voiced: $\beta = -.455, p < .01$; Emphatic: $\beta = -.443, p < .05$), but only in voiced stops for F0 ($\beta = .508, p < .01$) and SCG of burst ($\beta = -.241, p < .01$). These results indicate that categorization accuracy for the three stops varied differently when listeners used cues to identify coronal stops.

No pattern of categorical perception was observed for SCG of burst and F0 (see Figure 7C–D). Coefficients for these cues were positive indicating that accuracy increased as frequency values increased. This was consistent with the fact that listeners had higher accuracy rate for voiceless [t] and emphatic [t] than for voiced [d]. The pattern was different for F1. The coefficient for F1 was negative indicating that more accuracy was achieved at lower frequencies of the first formant. An interaction with stop category revealed that the tendency was the opposite for emphatic stops: they had higher accuracy rate as F1 frequency was higher. In addition, listeners showed a categorical response to F1 when they were to distinguish the contrast between emphatic [t] and voiced [d] (see Figure 7E). The categorical boundary was between 600 and 650 Hz; tokens with higher F1 were typically identified as emphatic. Interestingly, identification accuracy of voiceless stops was not sensitive to the changes in F1 values. Higher F1 values were associated with voiceless [t] in the whole range of frequencies, suggesting listeners were biased toward one of the ambiguous categories.

4.4 Summary and discussion

To sum up, the results of the perception experiment suggest that the overlap on the VOT scale does not lead to complete neutralization of the three categories of coronal stops. When VOT values fall into the ambiguous range, the contrast is still maintained in spectral cues. Listeners can use both noisy and vocalic portions of stops to identify such tokens. Performance was predictably worse in the noise condition; however, consonantal cues were sufficient for categorization of voiceless plain [t] and emphatic [tˤ] stops. Successful identification may be due to differences in spectral cues during the release and significantly shorter VOT of emphatic [tˤ]. In addition, presence of formant frequencies in a stimulus primarily affects performance in the emphatic [tˤ], as the contrast in emphasis is largely maintained on the adjacent vowel. Better performance on voiced [d] in the vocalic condition may also be due to a significant effect of F0 and F1 on categorization of voicing in coronal stops. The findings suggest that listeners use a complex pattern of categorization, which includes some compensation for phonological context, that is, preference toward a particular phonological category of a stop with a short-lag VOT.

5 Modeling cue integration and compensation for voice and emphasis

5.1 Background: Current models of perception

Results of the acoustic study and perception/identification study were used to create a computational model that could integrate acoustic cues for three categories of stops along the two phonological dimensions and account for phonological compensation observed in human listeners. Several models have been proposed to explain mechanisms of cue integration and parsing. The *Fuzzy Logical Model of Perception* (FLMP, Massaro & Oden, 1980; Oden & Massaro, 1978) is based on the assumption that listeners use unweighted integrated cues for "quick and dirty" analysis. The model emphasizes a holistic approach to speech recognition and assumes listeners compare the information from the acoustic signal with the characteristics of prototypes stored in long-term memory. The prototype that matches the information from an acoustic signal best is then used for categorization. Oden and Massaro (1978) successfully tested the model to discriminate place and voicing of the English stops /b, p, d, t/ based on information from a limited number of cues: F2-F3 for place of articulation and VOT for stop voicing. Massaro and Oden (1980) demonstrated that integration of cues occurred as an operation of "measuring" each acoustic feature against its value in a prototype sound. The model can potentially account for compensation as it includes so-called "modifiers" that evaluate a token as being distant or close to the prototype, but the prototype approach to categorization and similar invariance theories (e.g., Stevens & Keyser, 2010) do not typically require cue integration or compensation for context. Further studies (e.g., McMurray & Jongman, 2011) showed that the uncompensated model was not very successful in situations that required integration of multiple cues in a variable context, for example, categorization of English fricatives.

The mechanism of evaluation of each cue to reflect its relative importance in identifying a category was discussed in detail in the *Normal A Posteriori Probability* model (NAPP) (Nearey, 1990, 1997). Under this theory, perception was modeled as a sort of weighted sum—the evidence for the category from each cue weighted by the reliability of that cue. The computational mechanism for the model is based on logistic regression that incorporates simple effects of vowel and consonant as well as terms for interactions of different cues. The model was developed to account for recognition of CV or VC diphones, for example, Vt/d (Nearey, 1990, 1997) or s/ʃV (Nearey, 1992). In each of these cases, the same cue contributes to categorization of both segments in a diphone. For instance, vowel duration in a Vt/d diphone is a *primary cue* to voicing of a stop and a *secondary*

cue to vowel height, as low vowels are longer than non-low vowels. Similarly, F1 is a primary cue to vowel height and a secondary cue to stop voicing. The model accounted for compensation in categorization. Vowel and consonant choices were not independent of each other: listeners not only used primary and secondary cues to categorize vowels and consonants, but also used vocalic cues to predict a consonant and vice versa.

The model also accounted for cases of coarticulation in a CV diphone. For example, categorization of the syllables /si, su, ʃi, ʃu/ in English requires compensation for lip rounding, which lowers fricative mean frequency (center of gravity), but this gestural overlap is unidirectional in phonology. The vowel affects the fricative, but the fricative does not affect the vowel. Nearey (1992) argues that recognition of the segments in a diphone does not have to be hierarchical, as the effect of coarticulation in this case is bidirectional: speakers of English typically round the palatal fricative. This additional source of lip rounding affects the vowel articulation: not only the /s-ʃ/ boundary is lower before /u/, but also the /i-u/ boundary is lower before /ʃ/. The model correctly accounts for it by allowing each cue to simultaneously predict a vowel and a consonant.

The *Hierarchical Categorization* model (HICAT) by Smits (2001) builds upon NAPP by incorporating the principle of hierarchical dependency when categorization of one segment depends on the results of categorization of the other segment. Similar to NAPP, HICAT utilizes the computational mechanism of logistic regression, but it requires the terms enter the model in a hierarchical fashion. The model was used to account for recognition of the syllables /si, sy, ʃi, ʃy/ in Dutch. Unlike in English, the palatal fricative /ʃ/ in Dutch is not rounded in production. Therefore, rounding can spread only from the vowel but not *vice versa*. Successful categorization of a syllable in this case largely depends on a vowel. According to Smits (2001, p.1113), a hierarchical model is necessary when "coarticulation is high" <and> . . . "performance seriously degrades." Although both NAPP and HICAT would demonstrate similar performance in relatively simple cases, HICAT has an important advantage. It does not require a syllable/diphone as a recognition unit. Hierarchical processing can be viewed as a proxy of on-line processing of acoustic cues in spoken word recognition (McMurray et al., 2003; McMurray et al., 2009). This condition is particularly important when complexity of contextual factors becomes a limitation factor for NAPP.

As the task of segment recognition is performed not only in monosyllables but also in polysyllabic words, it must include compensation for cues not only in adjacent segments but also in the segments in the following syllable. The *Computing Cues Relative to Expectations* (C-CuRE) model demonstrates how variation in the acoustic signal can be reduced by attributing portions of the variation to context (Cole et al., 2010; McMurray et al., 2011). Similar to HICAT, C-CuRE utilizes the computational mechanism of hierarchical regression, but it expands the regression model to deal with additional sources of variation, for example, talker variation or V-to-V coarticulation. The model was tested to recover underlying categories of the vowels [ɛ] and [ʌ] in disyllabic words by removing contextual variance of talker, adjacent consonant, and following vowel from overlapping distributions of raw F1 and F2 values. Step-by-step removal of all contextual effects resulted in well-separated distribution of the two vowels along both dimensions: height and backness.

McMurray and Jongman (2011) showed that C-CuRE could accurately explain compensation for variation in cues in English fricatives and categorized fricatives similar to human listeners. The model coped with a wide variety of cues (the authors reported 24 cues, each being a significant predictor of a fricative category) that were mapped on three phonological categories to distinguish English fricatives /f, v, θ, ð, s, z, ʃ, ʒ/: place, voice, and sibilance. Although C-CuRE did not have 100% performance rate, it yielded an accuracy level similar to listeners and revealed the same error pattern shown by human listeners. When the model's performance was compared with the performance of FLMP and NAPP, C-CuRE substantially outperformed the two other models, especially when dealing with contextual factors. The model was significantly more accurate in predicting fricative categories as a function of talker and vowel. However, the model requires rich context to

predict correct categorization of a segment. It is not particularly clear how the model accounts for categorization when contextual information is insufficient.

All the models discussed above utilize the principle of cue integration, but they differ in how they treat relevance and reliability of individual cues. While FLPM represents integration of all available cues as a sum, NAPP (as well as HICAT, which departs from NAPP in that it uses hierarchical cue processing) distinguishes between primary and secondary cues by assuming the former are more important for categorization. Secondary cues can, nevertheless, also be relevant in ambiguous cases or serve as so-called "fudge factors" that boost performance (Nearey, 1997, p. 3248). Integration of cues and their relevance in these models is category-dependent, that is, listeners have to know the intended category in order to calculate relevance of each cue (see Toscano & McMurray, 2010, for an in-depth discussion of the problem). In contrast, C-CuRE assumes that cue evaluation mechanism is based on listeners' expectations, which emerge as a result of learning.

Toscano and McMurray (2010) developed a computational approach to evaluate weights for each cue from production data. Weights depend on reliability of a cue, which is the function of mean difference and the inverse function of variance. Smaller difference between distribution means and greater variance (hence, greater overlap between the categories) make a cue less reliable. Greater difference between the means and smaller variance, in contrast, result in little or no overlap between two distinct categories and make a cue a reliable predictor of a category. The algorithm proposed in Toscano and McMurray (2010) was tested to learn to discriminate two voicing categories in English word-initial stops using two cues: *VOT* (a strong, primary cue) and *duration of the following vowel* (a weak, secondary cue). The model with weighted cues correctly predicted effects of both cues and the trading relation between the two cues.

Further research has demonstrated that cue evaluation is an on-line process in which integration occurs relatively late because some cues, especially temporal ones, are simply not available at the same time (McMurray & Jongman, 2011). For example, listeners use vowel duration as secondary cue to voicing to adjust their interpretation of VOT in the ambiguous cases. Listeners are more likely to interpret the same positive VOT value as long-lag and thus categorize an English stop as voiceless before a shorter vowel, and as short-lag, that is, categorically voiced before a longer vowel (Allen & Miller, 1999). However, duration of vowel can only be assessed after assessment of VOT, so each parameter must be stored as an independent value in short-term memory (Toscano & McMurray, 2012). As a result, the effect of a secondary cue can change if listeners have enough information for categorization from other cues that covary with VOT (p. 1296).

5.2 Computational model of cue integration and phonological compensation

The primary goal of the study was to propose the model of cue integration and phonological compensation that can account for integration of cues to the two phonological contrasts. In line with the previous studies (Cole et al., 2010; McMurray & Jongman, 2011; Oden & Massaro, 1978; Nearey, 1990; Smits, 2001), modeling was achieved by using a logistic regression as a tool to analyze phoneme categorization. A logistic regression model evaluates the probability of a particular response via the logistic function, which is a monotonically increasing function with asymptotes at 0 and 1. For example, probability of realization of a stop as voiced can be evaluated using a binary dependent variable as shown in equation 1.

$$\mathrm{P}\left(voice\right) \ = \frac{1}{1+e^{-\left(\beta_0+\beta_1 x_1+\beta_2 x_2\ldots+\beta_n x_n\right)}} \tag{1}$$

Here, x_1 is a numerical value of a cue, and β is a weighted estimate of the cue.

Logistic regression makes a prediction about a category by linearly weighting and combining each cue and then converting them to a probability. These estimated weights typically reflect two things. First, they serve as a sort of scaling on the cues, converting them to a more scale-free metric (e.g., a cue that scales in milliseconds will have a very different range, than a formant frequency that scales in Hertz). In this case, the cues were converted to z-scores so that there was no scaling problem to solve. Second, and more importantly, the weight determines reliability of a cue as a function of how well the category can be predicted from the cue. The model is trained by giving it the correct label for each token, along with all the cue values. It then adjusts the weights to maximize its accuracy. It is important to test the model on different data than the model was trained on in order not to "overfit" the data. Thus, the regression model was trained on the cases from the first trial excluding tokens used in the perception experiment.

Probabilities of three categories can be evaluated using multinomial logistic regression, which combines regression parameters of several categories. In this case, the model to distinguish [d] and [ṭ] from [t] will have two regression parameters, one for each category except the reference category. Exponential of each category is shown in equation 2.

$$
\begin{aligned}
L(d) &= e^{(\beta_{d0} + \beta_{d1}x_{d1} + \cdot \beta_{d2}x_{d2} \ldots + \cdot \beta_{dn}x_{dn})} \\
L(\underline{t}) &= e^{(\beta_{t0} + \beta_{t1}x_{t1} + \cdot \beta_{t2}x_{t2} \ldots + \cdot \beta_{tn}x_{tn})}
\end{aligned}
\tag{2}
$$

They are combined to evaluate probability of any category:

$$
P(d) = \frac{L(d)}{1 + L(d) + L(\underline{t})}
\tag{3}
$$

Then the probability of the reference category is

$$
P(t) = \frac{1}{1 + L(d) + L(\underline{t})}
\tag{4}
$$

When two categories are identified in a binary logistic model, each cue receives only one weight as an estimate of reliability. In a multinomial model, weights can differ depending on the reference category; however, the difference in weights represents how each category is different with respect to a reference category. Therefore, modeling cue integration must include some kind of justification of the choice of the reference category. In case of Arabic coronals, voiceless [t] was chosen as a reference category because it is the unmarked category among the three stops. It is minimally different from [d] in voicing ([d] is [+voice] while [t] is [-voice]) and from [ṭ] in emphasis ([ṭ] is [+emphatic while [t] is [-emphatic]).

Although this coding makes it possible to put the stops in the three models on the same scale in a singular dimension, it does not represent contrast-specific weighting of a cue. Recall that [d] is different in voicing not only from the reference category [t], but also from the emphatic category [ṭ], which is also [–voice]. Similarly, [d] is also different from [ṭ] as it is [-emphatic]. Contrast-specific weighting should include evaluation of each contrast in a separate orthogonal dimension. Thus, in order to compare all models directly, the categories in the orthogonal model were converted to be compared to the reference category on a single scale.

5.3 Models of cue integration: Overall data

Three models that represent common approaches to categorization and cue weighting were probed. Model 1 was designed following the logic of FLMP (Massaro & Oden, 1980; Oden & Massaro,

Table 10. Summary of indices produced by the three categorization models.

Stop	Parameter	Model 1	Model 2	Model 3
		Unweighted, unadjusted	Weighted, adjusted for context	Weighted, adjusted for contrasts
Voiceless [t]	Mean (*SD*)	.881 (.82)	.965 (.76)	1.007 (.36)
	Estimate (*SE*)	.880 (.19)***	.964 (.14)***	1.007 (.08)***
Voiced [d]	Mean (*SD*)	−.328 (.87)	−.839 (.57)	−1.002 (.79)
	Estimate (*SE*)	−1.208 (.28)***	−1.803 (.22)***	−2.009 (.16)***
Emphatic [t]	Mean (*SD*)	−.707 (.31)	−.161 (.53)	−.007 (.14)
	Estimate (*SE*)	−1.587 (.27)***	−1.126 (.21)***	−1.014 (.10)***
	Log Likelihood	1842.9	1349.1	585.4
	Cues	VOT, F2	All 5	All 5

Note: significance level: *** – $p < .001$.

1978) and invariance theories (Stevens & Keyser, 2010). It used two major cues to voicing and emphasis—VOT and F2—that were unweighted for context and phonological contrasts. The model predicts that listeners should use salient and invariant acoustic cues and compare their values with prototypical values for "quick and dirty" categorization. While this may be sufficient for categorization of unambiguous stops, it is not clear how the model deals with potentially ambiguous tokens.

Model 2 was designed in line with perception models that use weighted cues and compensate for context (NAPP, Nearey, 1990; HICAT, Smits, 2001; C-CuRE, Cole et al., 2010; McMurray & Jongman, 2011). The model predicts that listeners should adjust cues by assigning weights and removing talker and vowel contextual variance. But cues in this model are not weighted for particular phonological contrasts. They represent overall balanced reliability of cues to identify a set of categories. Also, the model is sensitive to richness of context, and it may not reliably predict ambiguous tokens in the absence of sufficient contextual information.

Model 3 was designed to utilize the principle of compensation for phonological contrasts. It used cues that were weighted for each phonological contrast separately. Although the model is expected to show similar accuracy of performance with Model 2 to categorize unambiguous stops, contrast-specific weighting should be beneficial for categorization of stops with potentially ambiguous VOT.

First, the three models were tested for their capacity to fit the production data. The model summary is shown in Table 10; the effects of the stop category on accuracy score are presented in Figure 8. Stop categories were represented in a continuous fashion using indices that summarized acoustic information from the cues. These indices were unweighted (Model 1) or weighted (Models 2 and 3) sums of normalized values (z-scores) of the cues, as shown in equation (1). Each index showed how likely the token is to be identified as an intended category. When normalized indices were fit to the mixed-effects linear model with stop category as a fixed factor, and talker and item as random factors, the analysis revealed significant effects of stop category for all models. Each model distinguished the voiceless, voiced, and emphatic categories and differences between them.

5.3.1 Model 1: Major unweighted cues. Model 1 asked if a sum of several unweighted cues is sufficient to discriminate the three categories of coronal stops. Two major cues that were found best predictors for voice and emphasis–VOT and F2– were used in this model. When the model was fit to the data to predict the underlying category of a coronal stop, it achieved a relatively good fit, Log likelihood = 1842.9, $\chi^2(2) = 769.8, p < .0001$, averaging 57% correct. The model correctly predicted voiceless stops (87%), but it performed poorly to predict the voiced (36%) and emphatic (43%) categories.
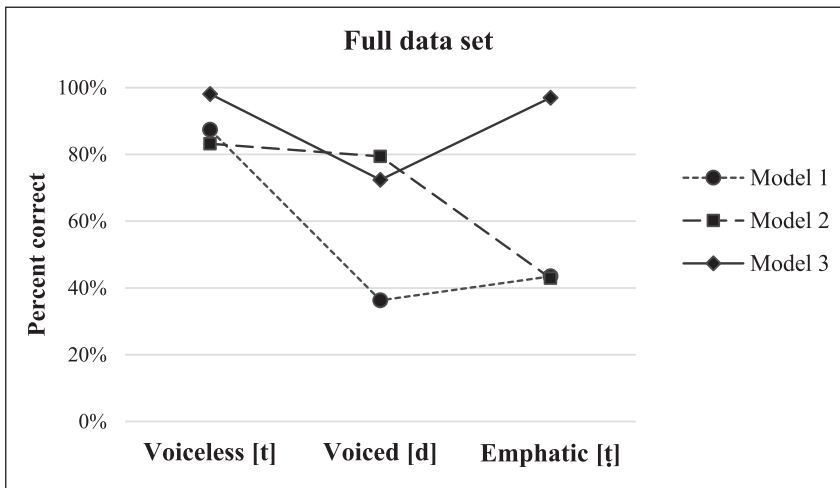
Figure 8. Performance of the three multinomial models to predict three coronal stops in Qatari Arabic across the complete data pool.

5.3.2 Model 2: Cues compensated for context. Model 2 used a sum of cues adjusted for context, such as talker and vowel in line with C-CuRE. Adjustment of each cue was performed in a hierarchical regression model. After dummy codes for talker (7) and vowel (2) were entered the model, the cues were recoded as standardized residuals. Although variables in a regression model are typically weighted as parameter estimates, these weights were not explicitly specified or evaluated for two different phonological contrasts. They were treated as a measure for each cue to discriminate the three categories in the same dimension. Model 2 showed better performance than Model 1. It achieved a good fit, Log likelihood = 1349.1, $\chi^2(2) = 1124.7$, $p < .0001$, averaging 71% correct. Similar to Model 1, it performed very well on voiceless (83%) and showed less accuracy on emphatic stops (43%), but it showed a significant improvement in categorization of voiced stops (80%).

5.3.3 Model 3: Cues compensated for phonological contrasts. Model 3 was based on the assumption that listeners can assign contrast-specific weights to the same acoustic cues and compensate for phonological contrasts. Following McMurray and Jongman (2011), the two logistic regression models were trained to predict the contrasts in each stop from the five cues. To distinguish voiced [d] from voiceless stops, the coefficients for VOT, SCG of burst, F0, and F1 were negative, meaning that voiced stops had negative VOT, lower spectral mean, lower fundamental frequency, and lower frequency of the first formant than voiceless stops. The coefficient for F2 was positive, which was consistent with the fact that voiced stops had higher F2 frequencies. To distinguish emphatic [ṭ] from non-emphatic stops, the coefficients for VOT, SCG of burst, F0, and F2 were negative, meaning that emphatic stops had shorter VOT, lower spectral mean, lower fundamental frequency, and lower frequency of the second formant than non-emphatic stops. The coefficient for F1 was positive indicating that emphatic stops had higher F1 frequency.

When the model with cues adjusted for contrast was fit to the total pool of data to predict the underlying category of a coronal stop, it achieved a good fit, Log likelihood = − 585.4, $\chi^2(2) = 1881.9$, $p < .0001$, averaging 89% correct. Unlike Model 1 and Model 2, this model performed very well on all three categories confusing mostly voiced and voiceless categories.
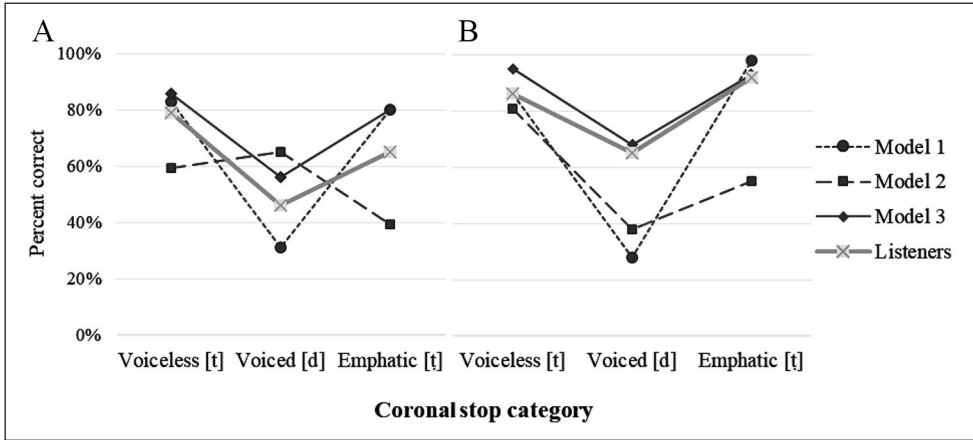
Figure 9. Performance of the three models and human listeners to predict coronal stops in Qatari Arabic overlapping on the VOT scale in A) Noise and B) Vocalic conditions.

5.4 Resolving ambiguity in stops overlapping on the VOT scale

Comparison of the models presented in the previous section showed that each of the three models distinguished three categories of coronal stop, although their predictive power was different. As expected, Models 1 and 2 with weighted cues outperformed Model 1 with unweighted cues indicating that adjusting cues for context and phonological contrast can improve categorization. At the next step, the three models were tested for their ability to resolve ambiguity in coronal stops overlapping on the VOT scale similar to human listeners. 360 tokens ($n = 120$ per category) were selected for the analysis. They were voiced [d]s and voiceless [t]s produced with VOT within the range between 0 and 40 ms in Experiment 1. In addition, emphatic stops were randomly selected from the production pool to match the number of non-emphatic stops. The pool of ambiguous tokens included all stops used in Experiment II to ensure the link between performance of the computational model and human listeners' performance. The analysis of models replicated the design of the perception experiment. Each model was tested twice: first, with consonantal cues only (Figure 9A), which was analogous to the Noise condition in Experiment II, and then with both consonantal and vocalic cues (Figure 9B) similar to the Vocalic condition in Experiment II.

When *Model 1* was fit to the set of ambiguous data in the Noise condition, it achieved a good fit, Log likelihood: 389.8; $\chi^2(2) = 353.4, p < .0001$, but it correctly identified only 44% of tokens. The model correctly predicted 83% of [t]s, 80% of [t]s, but only 31% of voiced [d]s were predicted correctly. 44% of them were confused with emphatic stops, and 25% with voiceless stops. The model achieved a slightly better fit when vocalic cues were added, Log likelihood: 325.9; $\chi^2(2) = 418.7, p < .0001$. It correctly predicted 73% of tokens: 86% of [t]s and 98% of [t]s, but surprisingly it showed less accuracy categorizing voiced [d]s (28%).

When *Model 2* was fit to the set of ambiguous data in the Noise condition, it achieved a good fit, Log likelihood: –593.3; $\chi^2(2) = 126.9, p < .0001$, and outperformed Model 1. It correctly identified 54% of tokens, and, importantly, it discriminated all three categories. It correctly predicted 59% of [t]s, 65% of voiced [d]s and 39% of emphatic [[t]s. The model's performance improved in the Vocalic condition, Log likelihood: –495.5, $\chi^2(2) = 231.1, p < .0001$, with identification rate averaging 59%. The model correctly predicted 81% of [t]s, 38% of voiced [d]s and 55% of emphatic [t]s. Most confusions (62%) happened between the voiced and emphatic categories.

Model 3 also discriminated all three categories. It achieved a good fit, Log likelihood: 331.9, $\chi^2(2) = 355.3, p < .0001$, and outperformed Models 1 and 2. With consonantal cues, it correctly

identified 72% of tokens. The model correctly predicted 86% of [t]s but performed less accurately on voiced [d]s (50%) and emphatic [ṭ]s (77%). Most confusions occurred between the voiced and voiceless categories. In the vocalic condition, the performance improved considerably, Log likelihood: –98.8, $\chi^2(2) = 570.4$, $p < .0001$, achieving 92% correct identifications. The model correctly predicted 93% of [t]s, 90% of voiced [d]s and 94% of emphatic [ṭ]s.

5.5 Summary

The results showed that each model distinguished between the three categories of coronal stops and could predict ambiguous categories in some conditions similar to human listeners. But only Model 3 revealed performance similar to performance of human listeners in all conditions. The findings suggest that weighting cues with respect to both phonological dimensions is beneficial in a task that requires resolving ambiguity. Model 3 demonstrated overall accuracy similar to listeners' accuracy, as well as the same types or error and confusion patterns as human listeners. It failed to predict ambiguous [d]s and confused them mostly with voiceless [t]s.

Some results were surprising. For example, the model with unweighted and unadjusted cues (Model 1) could correctly predict voiceless [t]s and emphatic [ṭ]s in the Noise condition similar to Model 3 and listeners. These findings suggest that VOT is indeed the most salient invariant consonantal cue that can identify the three categories of stops. High accuracy rate for emphatic stops suggest that short-lag VOT is an important cue to emphasis in Qatari Arabic stops. Although the analysis of cues in section 3.3.2 suggested that VOT may not be associated with emphasis, listeners' performance indicated that this prediction was incorrect. Short-lag VOT was a very reliable predictor of the emphatic category in coronal stops. It is possible that the results of the analysis were confounded by the fact that the emphatic category falls *between* the voiced and voiceless categories on the VOT scale, which masked the effect of this cue in the acoustic analysis.

Despite high accuracy rate for voiceless [t] and [ṭ], Model 1 performed poorly on voiced [d]s with short-lag VOT, confusing them with voiceless categories that also had positive VOT values. Inadequate performance of the model in the vocalic condition suggests that the effect of F2 was insufficient. Model 2 showed a significantly better performance on average, but, unexpectedly, it showed a low accuracy rate on emphatic stops, especially in the vocalic condition, when only one vocalic context was used. These results seem to be counterintuitive because using additional cues should improve performance of the model. It is possible that Model 2 indeed requires rich context and cannot adequately deal with cases that involve insufficient contextual information.

6   General discussion and conclusion

The primary goal of the study was to determine how the phonological contrasts of voicing and emphasis in coronal stops in Qatari Arabic are distributed in the multidimensional acoustic space. To address this question, a corpus of production data was collected. The acoustic analysis of the cues showed that both contrasts were linked to essentially the same cues, and that every cue influenced contrast identity. Therefore, proper mapping of the cues on the three categories of stops—voiceless [t], voiced [d], and emphatic [ṭ]—requires some kind of weighting and contrast-specific compensation. Compensation for phonological contrast is particularly important for correct categorization of stops that overlap on the VOT scale.

Listeners' performance was used in this study to achieve a two-fold goal. First, the results of the perception study revealed that listeners are capable to resolve ambiguity even if the contrast is partially neutralized. Evaluation of the production data showed that the overlap in the potentially ambiguous range between 0 and 40 ms does not result in complete neutralization of the two contrasts. VOT values

in the three categories of coronal stops remained distinct, although the differences between voiced, voiceless and voiceless emphatic categories became very small. For example, short-lag positive VOT values in voiced [d]s were only 4 ms longer than short-lag positive VOT values in emphatic [ṭ]s—a difference that may not be detected in perception (Abramson & Lisker, 1970). The findings suggest that listeners were able to discriminate two quasi-categories inside the short-lag range. Stops with longer VOT (35-40 ms) were categorized as a voiceless "aspirated" [t] but stops with shorter VOT (10-30 ms) consisting of tokens of voiced [d] that lacked prevoicing and of tokens of voiceless emphatic [ṭ] were categorized as an "unaspirated" category. Most importantly, categorization of "unaspirated" stops revealed a preference toward the emphatic category.

The response patterns observed in the study suggest that listeners may use compensation for phonological contrast, showing preference toward one of the categories when cue values are largely neutralized. For example, short-lag VOT was more often associated with emphatic [ṭ] than with unvoiced [d]. In the absence of other contextual factors, this preference must be derived from phonological knowledge. Speakers of Qatari Arabic use short-lag VOT as a cue to emphasis and tend to perceive voiceless unaspirated stops as emphatic. Longer VOT, in contrast, is associated with voiceless non-emphatic [t]. A similar pattern was observed in F2 as well. Listeners typically associated higher F2 values with non-emphatic stops, but they did it more often for voiceless [t] than for voiced [d].

Second, listeners' accuracy rate on the set of ambiguous tokens was used to evaluate the performance of the computational models. To investigate how listeners integrate multiple cues for two orthogonal contrasts in stops, three models of categorization were evaluated and compared to listener performance. The models were in line with three common assumptions of theoretical models of categorization and were used to test the following predictions: a) that a few uncompensated cues may be sufficient for categorization, b) that weighted cues compensated for context are sufficient, and c) that cues may as well be compensated for phonological contrasts.

The findings suggest that compensation for phonological contrasts observed in listeners' categorization tasks can be encoded as weighting of cues in computational models. Such weighting seems to be different from weighting previously reported in the literature (e.g., Nearey, 1990, 1992). Cue weights used in Model 3 in the current study are not just the result of adjustment for the immediate phonetic context. They are also the result of adjustment for the phonological contrasts in the language. Listeners showed preference for a particular category of a stop in the same phonetic context, assigning potentially ambiguous tokens with shorter VOT values to the emphatic category.

Previous studies (e.g., Cole et al., 2010; McMurray et al., 2011) reported that the models with weighted cues performed similar to human subjects. Although the current study revealed a similar pattern, some results deviate from the patterns observed in literature. For example, McMurray and Jongman (2011) report that performance of categorization models did not exceed performance of human subjects. In the current study, the model with cues compensated for phonological contrasts (Model 3) showed a higher accuracy rate than listeners. At least two explanations are possible. First, the high accuracy rate may be the result of over-saturation of the model. Recall that the weights were estimated by providing the model with correct labels for the contrasts, but listeners had to identify the type of contrast during the task before assigning the relevant weights. Although weights were estimated on a subset of data and were applied to the rest of tokens (the technique typically used when model training is required, see McMurray & Jongman, 2011, or Toscano & McMurray, 2010, for details), this procedure was probably not sufficient to desaturate the model. High performance rate may also be viewed as a limitation of the model and the method. It suggests that the actual human listeners' performance probably includes other factors that were not represented in the model. But in general, the results of the performance of the model can be treated as plausible because Model 3, unlike other models, also demonstrated the error pattern similar to human listeners. The lower performance of the C-CuRE model in McMurray and Jongman (2011) could be due to the differences between stops and fricatives. Fricatives in general are more difficult in perception (see Jongman

et al., 2011, among others) and require a larger number of cues than stops (23 cues were used to test C-CuRE, none of which was invariant). It is possible, that stops were better identified by a computational model in the Noise condition in this study because Model 3 used salient cues (e.g., VOT, an invariant cue to voicing) that significantly contributed to categorization.

Next, the findings suggest that compensation for context and compensation for phonological contrast might be different tasks for listeners. Both models that used cue weighting—Model 2 and Model 3—showed good performance on coronal stops within the ambiguous VOT range, and they were able to differentiate between the three categories of stops approaching the results of human listeners. However, the two models had different predictions about the confusion patterns. Model 2 (compensation for context) predicted confusion between voiced [d] and emphatic [t] due to similar weights of the spectral cues for the two categories of stops. Model 3 (phonological compensation) predicted confusion between voiced [d] and voiceless [t]. In this task, weighting all available cues facilitated discrimination of the categories, but adjustment of cues for phonological contrast facilitated more accurate identification of each category. The results of identification test by human listeners support Model 3, as listeners were able to differentiate all three categories of stops that overlap on the VOT scale and demonstrated a similar confusion pattern performing poorly on voiced [d]s and confusing it mostly with voiceless stops.

Although the model with phonological compensation (Model 3) showed more accurate performance than the model with compensation for context (Model 2), the results of the current study are not in conflict with predictions of C-CuRE. The difference between the two approaches lies in the scenarios where each model achieves its best results. The C-CuRE approach is aimed at dealing with rich contextual information. It ensures that listeners filter out contextual variation and adjust cues taking into account any contextual factors that can variably modify the acoustic signal, for example, speaker's gender, category of adjacent segment(s), etc. Weights in C-CuRE are parameters that emerge as a measure to evaluate relevance of cues to discriminate contrastive categories. They indicate to what extent the cue is modified by context and, hence, to what extent it can be reliable in this context.

The approach proposed in this paper, in contrast, aims to deal with situations when context is scarce, the categories are ambiguous, and the phonological contrast is partially neutralized. The model with phonological compensation predicts that listeners would still weight cues, but such weighting can be viewed as their expectations that originate from listeners' phonological knowledge. This type of weighting may work as a "shortcut" in processing of an acoustic signal when contextual information is insufficient for categorization. This approach may be beneficial for listeners as it allows them to categorize tokens successfully in situations when categories overlap.

Adjusting cues for phonological contrast is not incompatible with predictions of C-CuRE or HICAT. Listeners must identify speakers' identity/gender and the categories of all adjacent segments "simultaneously and interactively" (McMurray & Jongman, 2011, p. 240). Such identifications occur in parallel, but since listeners have to wait until all cues become available, they "may be able to revise their initial decisions" (ibid.). Weighting and adjusting cues for phonological contrast can be part of this process. When more than one competitor is activated in parallel processing, listeners may facilitate the decision process by excluding some of the candidates earlier than other candidates because they find them less relevant. In addition, parsing the acoustic signal to categorize a segment is a process that requires adjustments at the intermediate stage. Both C-CuRE and HICAT predict that partial parsing of cues will help to make a preliminary decision about an adjacent segment, and then this contextual information will help to identify the segment (McMurray & Jongman, 2011; Smits, 2001). Phonological compensation is not in conflict with this prediction. It simply adds an additional source of information when actual phonetic context is not rich enough.

To conclude, cue integration is an essential part of speech recognition, and any successful parsing model must account for compensation and weighting of cues. Rich literature on the subject focuses primarily on various mechanisms that explain compensation of acoustic cues for phonetic

context. Less is known about the role of phonological context in cue weighting. The current study is an attempt to provide at least some answers to this question. The categorization model tested in the study uses cues weighted in relation to two phonological contrasts: voicing and emphasis. These contrasts in coronal stops are not symmetrical, and they and are linked essentially to the same set of acoustic cues. Thus, relevance of each cue to a particular phonological contrast is a vital problem for the parsing mechanism.

The findings suggest that in addition to adjustment and compensation for phonetic context, listeners may adjust cues based on their expectations derived from knowledge of the phonological system of a language. The model with cues separately weighted for different phonological contrast could identify segments more accurately and demonstrated the confusion pattern similar to the pattern observed in human subjects. Some findings of the study also suggest that cue weighting and decisions about phonological contrasts can be made in order that reflects a hierarchy of phonological contrasts. However, particular details of this process and the directionality of such hierarchies require additional research.

ORCID iD

Vladimir Kulikov https://orcid.org/0000-0001-9787-4801

References

Abramson, A. S., & Lisker, L. (1970). Discrimination along the voicing continuum: Cross-language tests. In *Proceedings of the Sixth International Congress of Phonetic Sciences, Prague. 1967* (pp. 569–573). Academia Publishing House of the Czechoslovak Academy of Science.

Allen, J. S., & Miller, J. L. (1999). Effects of syllable initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America*, *106*(4), 2031–2039. https://doi.org/10.1121/1.427949

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boersma, P., & Weenink, D. (2018). PRAAT: Doing phonetics by computer. (Version 6.037). http://www.praat.org/

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd edition). Erlbaum.

Cole, J. S., Linebaugh, G., Munson, C., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*(2), 167–184. https://doi.org/ 10.1016/j.wocn.2009.08.004

Darcy, I., Ramus, F., Christophe, A., Kinzler, K., & Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. In F. Kügler, C. Féry, R. van de Vijver (Eds.), *Variation and gradience in phonetics and phonology* (pp. 265–310). Mouton De Gruyter.

Feghali, H. J. (2008). *Gulf Arabic: The dialects of Kuwait, Bahrain, Qatar, UAE, and Oman*. Dunwoody Press.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics*, *36*(4), 359–368. https://doi.org/10.3758/BF03202790

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception and Psychophysics*, *68*(2), 161–177. https://doi.org/10.3758/BF03193666

Fowler, C. A., & Brown, J. M. (2000). Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Perception and Psychophysics*, *62*(1), 21–32. https://doi.org/10.3758/BF03212058

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 144–158. https://doi.org/10.1037/0096-1523.22.1.144

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics*, *65*(4), 575–590. https://doi.org/10.3758/BF03194584

Jongman, A., Herd, W., Al-Masri, M., Sereno, J., & Combest, S. (2011). Acoustics and perception of emphasis in Urban Jordanian Arabic. *Journal of Phonetics*, *39*(1), 85–95. https://doi.org/10.1016/j.wocn.2010.11.007

Khattab, G., Al-Tamimi, F., & Heselwood, B. (2006). Acoustic and auditory differences in the /t/-/ṭ/ opposition in male and female speakers of Jordanian Arabic. In S. Boudelaa (Ed.), *Perspectives on Arabic Linguistics XVI* (pp. 131–160). John Benjamins.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. Language, *70*(3), 419–454. https://doi.org/10.2307/416481

Kulikov, V. (2016). Voicing in Qatari Arabic: Evidence for prevoicing and aspiration. *Qatar Foundation Annual Research Conference Proceedings*, *2016*(1). https://doi.org/10.5339/qfarc.2016.SSHAPP2330

Kulikov, V. (2020). Laryngeal contrast in Qatari Arabic: Effect of speaking rate on VOT. *Phonetica*, *77*(3), 163–185. https://doi.org/10.1159/000497277

Kulikov, V., Mohsenzadeh, F., & Syam, R. M. (2020). Effect of emphasis spread on coronal stop articulation in Qatari Arabic. *Proceedings of the Linguistic Society of America*, *5*(1), 16–28. https://doi.org/10.3765/plsa.v5i1.4652

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(2), 1–26. https://doi.org/10.18637/jss.v082.i13

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Blackwell.

Lenth, R. (2020). Emmeans: Estimated Marginal Means, aka Least-Squares Means. (R package version 1.47). https://CRAN.R-project.org/package=emmeans

Lisker, L, & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, *29*(Pt 1), 3–11. https://doi.org/10.1177/002383098602900102

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). *Journal of the Acoustical Society of America*, *102*(2), 1134–1140. https://doi.org/10.1121/1.419865

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception and Psychophysics*, *60*(5), 602–619. https://doi.org/10.3758/bf03206049

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception and Psychophysics*, *28*(5), 407–412. https://doi.org/10.3758/bf03204884

Massaro, D. W., & Oden, G. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice*, vol. 3 (pp. 129–165). New York: Academic Press.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

McCarthy, J. J. (1994). The phonetics and phonology of Semitic pharyngeals. In P. A. Keating (Ed.), *Phonological structure and phonetic form: Papers in Laboratory Phonetics III* (pp. 191–233). Cambridge University Press.

McMurray, B., Tanenhaus, M. K., Aslin, R. N., & Spivey, M. J. (2003). Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access. *Journal of Psycholinguistic Research*, *32*(1), 77–97. https://doi.org/10.1023/a:1021937116271

McMurray, B., Tanenhaus, M., & Aslin, R. (2009). Within-category VOT affects recovery from ''lexical'' garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 65–91. https://doi.org/10.1016/j.jml.2008.07.002

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219–246. https://doi.org/10.1037/a0022325

McMurray, B., Cole, J. S., & Munson, C. (2011). Features as an emergent product of computing perceptual cues relative to expectations. In R. Ridouane & N. Clements (Eds.), *Where do features come from*? (pp. 197–236). John Benjamins.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, *43*(1–3), 106–115. https://doi.org/10.1159/000261764

Mitleb, F. (2009). Voice onset time of Jordanian Arabic stops. *Journal of the Acoustical Society of America*, *109*(5), 2474. https://doi.org/10.1121/1.4744787

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*(5), 2088–2113. https://doi.org/10.1121/1.397861

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, *18*(3), 347–373. https://doi.org/10.1016/S0095-4470(19)30379-1

Nearey, T. M. (1992). Context effects in a double-weak theory of speech perception. *Language and Speech*, *35*(1–2), 153–171. https://doi.org/10.1177/002383099203500213

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, *101*(6), 3241–3254. https://doi.org/10.1121/1.418290

Oden, G., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*(3), 172–191. https://doi.org/10.1037/0033-295X.85.3.172

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *Journal of the Acoustical Society of America*, *75*(1), 224–230. https://doi.org/10.1121/1.390399

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org

Repp, B. H. (1983). Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. *Speech Communication*, *2*(4), 341–361. https://doi.org/10.1016/0167-6393(83)90050-X

Smits, R. (2001). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1145–1162. https://doi.org/10.1037/0096-1523.27.5.1145

Stevens, K. N., & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics*, *38*, 10–19. https://doi.org/10.1016/j.wocn.2008.10.004

Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, *62*(2), 435–448.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. https://doi.org/10.1111/j.1551-6709.2009.01077.x

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception & Psychophysics*, *74*(6), 1284–1301. https://doi.org/10.3758/s13414-012-0306-z

van Alphen, P. M., & Smits, R. (2004). Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: The role of prevoicing. *Journal of Phonetics*, *32*(4), 455–491. https://doi.org/10.1016/j.wocn.2004.05.001

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology. Human Perception and Performance*, *36*(4), 1005–1015. https://doi.org/10.1037/a0018391

Watson, J. C. E. (2002). *The phonology and morphology of Arabic*. Oxford University Press.

Westbury, J. R. (1983). Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *Journal of the Acoustical Society of America*, *73*(4), 1322–1336. https://doi.org/10.1121/1.389236

Yeni-Komshian, G. H., Caramazza, A., & Preston, M. S. (1977). A study of voicing in Lebanese Arabic. *Journal of Phonetics*, *5*(1), 35–48.

Zawaydeh, B. A., & de Jong, K. J. (2002). Uvularization spread in Arabic. *Speech Prosody and Timing: Dynamic Aspects of Speech: IULC Working Papers in Linguistics*, *2*(2), 93–107.

Appendix 1

Table 4. Summary of the mixed-effects linear model examining human listeners' categorical response pattern in stops overlapping on a VOT scale. The reference category is identification of voiceless stop [t].

Predictor	Level	Estimate	Std. Error	t value	Pr (> |t|)
Fixed effects					
	(Intercept)	.79	.04	19.31	< 0.0001
Condition	Vocalic	.07	.05	1.43	0.163
Stop	Voiced	−.32	.03	−9.44	< 0.001
	Emphatic	−.14	.03	−4.07	< 0.001
Condition: Stop	Vocalic: Voiced	.11	.03	3.63	< 0.001
	Vocalic: Emphatic	.20	.03	6.63	< 0.0001
Random effects					
Item	(Intercept)	.02			
Subject	(Intercept)	.02			
	Voiced	.02			
	Emphatic	.02			
Residual		.15			

Table 5. Summary of the mixed-effects linear model examining human listeners' response pattern as a function of VOT binned by 5 ms. The reference category is voiceless stop [t].

Effect	Level	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)		.616	.123	4.991	< .0001
VOT		.004	.004	1.054	n.s.
Condition	Vocalic	.275	.175	1.575	n.s.
Stop	Voiced	.038	.161	.239	n.s.
	Emphatic	.257	.153	1.681	n.s.
VOT: Condition	Vocalic	−.006	.006	−1.155	n.s.
VOT: Stop	Voiced	−.017	.007	−2.514	.013
	Emphatic	−.019	.006	−3.418	.001
Condition: Stop	Vocalic: Voiced	−.194	.227	−.855	n.s.
	Vocalic: Emphatic	−.217	.216	−1.003	n.s.
VOT: Condition: Stop	Vocalic: Voiced	.021	.010	2.213	.028
	Vocalic: Emphatic	.021	.008	2.656	.008
Random effects					
Subject	(Intercept)	.012			
	Voiced	.017			
	Emphatic	.028			
Residual		.053			

Table 6. Summary of the mixed-effects linear model examining human listeners' response pattern as a function of F2 binned by 200 Hz. The reference category is voiceless stop [t].

Effect	Level	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)		.091	.189	.479	n.s.
F2		.0004	.0001	3.617	< .0001
Condition	Vocalic	.108	.051	2.093	n.s.
Stop	Voiced	.642	.236	2.720	.007
	Emphatic	.689	.265	2.595	.010
Stop : F2	Voiced	−.0005	.0001	−3.914	< .0001
	Emphatic	−.0004	.0002	−2.140	.033
Random effects					
Subject	(Intercept)	.014			
	Voiced	.032			
	Emphatic	.042			
Residual		.054			

Table 7. Summary of the mixed-effects linear model examining human listeners' response pattern as a function of SCG of burst binned by 500 Hz. The reference category is voiceless stop [t].

Effect	Level	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)		.741	.041	18.284	< .0001
SCG of burst		.00004	.00001	4.241	< .0001
Condition	Vocalic	.085	.057	1.482	n.s.
Stop	Voiced	−.241	.081	−2.993	.004
	Emphatic	.003	.076	1.681	n.s.
SCG: Condition	Vocalic	−.00001	.00001	−.568	n.s.
SCG: Stop	Voiced	−.00008	.00003	−2.346	.019
	Emphatic	−.00013	.00003	−3.724	< .001
Condition: Stop	Vocalic: Voiced	.099	.114	.873	n.s.
	Vocalic: Emphatic	.068	.108	.634	n.s.
SCG: Condition: Stop	Vocalic: Voiced	.00000	.00005	.083	n.s.
	Vocalic: Emphatic	.00012	.00005	2.433	.015
Random effects					
Subject	(Intercept)	.008			
	Voiced	.035			
	Emphatic	.026			
Residual		.037			

Table 8. Summary of the mixed-effects linear model examining human listeners' response pattern as a function of F0 binned by 5 Hz. The reference category is voiceless stop [t].

Effect	Level	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)		.403	.074	5.423	< .0001
F0		.002	.0003	5.744	< .0001
Condition	Vocalic	.131	.043	3.025	.006
Stop	Voiced	.508	.142	3.565	< .001
	Emphatic	−.093	.168	−.552	n.s.
F0 : Stop	Voiced	−.003	.0006	−5.564	< .0001
	Emphatic	.0003	.0007	.442	n.s.
Random effects					
Subject	(Intercept)	.015			
	Voiced	.017			
	Emphatic	.030			
Residual		.052			

Table 9. Summary of the mixed-effects linear model examining human listeners' response pattern as a function of F1 binned by 100 Hz. The reference category is voiceless stop [t].

Effect	Level	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)		1.0	.082	12.177	< .0001
F1		−.0003	.0001	−2.800	.005
Condition	Vocalic	.106	.043	2.465	.021
Stop	Voiced	−.455	.149	−3.054	.002
	Emphatic	−.443	.175	−2.538	.011
F1: Stop	Voiced	.0003	.0003	1.169	n.s.
	Emphatic	.0006	.0002	2.336	.020
Random effects					
Subject	(Intercept)	.008			
	Voiced	.029			
	Emphatic	.046			
Residual		.053			