Cue-weighting in the perception of intervocalic stop voicing in European Portuguese

Daniel Pape[a)]
Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, 3830-193 Aveiro, Portugal

Luis M. T. Jesus
Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and School of Health Sciences (ESSUA), University of Aveiro, 3830-193 Aveiro, Portugal

(Received 26 October 2012; revised 25 June 2014; accepted 3 July 2014)

This paper describes the perception of intervocalic stop voicing in European Portuguese (EP) stimuli without a stop burst, when varying three acoustic cues: *Vowel duration*, *stop duration*, and *voicing maintenance* during stop closure. Perceptual stimuli were generated using biomechanical modeling. First, a discrimination experiment was conducted to determine the listeners' perceptual sensitivity to the *voicing maintenance* cue. Second, an identification experiment was conducted to examine the effect and interaction of *vowel duration*, *stop duration*, and *voicing maintenance* during stop closure on the voiced/voiceless identification responses of EP listeners. The results of the discrimination test show that *voicing maintenance* differences have a significant effect as soon as they exceed a certain threshold. In the identification experiment, evidence was found that only the two factors *vowel duration* and *voicing maintenance* significantly influence the listeners' decisions, but not *stop duration*. The ratio between *stop duration* and *vowel duration* plays a major role in distinguishing stop voicing, but only for highly devoiced stimuli. It is shown that in stimuli without a stop burst, both *voicing maintenance*, as a major but not required cue, and *vowel duration* are important acoustic cues for stop voicing distinctions in EP. © *2014 Acoustical Society of America.*
[http://dx.doi.org/10.1121/1.4890639]

PACS number(s): 43.71.Es, 43.71.Bp [BRM] Pages: 1334–1343

I. INTRODUCTION

The extensive research that has been dedicated to the perception of stop voicing has shown that several acoustic cues (separately or combined) form the perceptual construct of a voicing distinction. In other words, human perception does not rely only on a single perceptual cue, but rather on a combination of different cues to guarantee a stable and robust perceptual outcome. One phonological contrast can be triggered by multiple acoustic cues, or by a combination of these cues. For the perception of stop voicing, voice onset time (VOT) is seen as one of the major acoustic cues in a variety of languages, the others being stop duration and adjacent vowel duration (Cuartero, 2002; Jessen, 1999; Lisker and Abramson, 1967; Luce and Charles-Luce, 1985; Viana, 1984). VOT is most important in prosodically strong positions, like word-initial position and stress foot-initial position.

For stops in medial position, the traditional definition of a voiced stop is one with *voicing during closure* [see, for example, Lisker (1986) for English, Jessen (1999) for German]. English intervocalic stops usually have voicing during closure, which strongly influences perception (Francis *et al.*, 2000; Lisker, 1986); however, other acoustic cues like stop duration (Francis *et al.*, 2000) and preceding vowel duration (Raphael, 1972) strongly influence voicing perception. Adding to this complexity, Li *et al.* (2010), among others, showed that a number of other acoustic cues ($f0$, $F1$, $F2$, burst amplitude) may also play a role as perceptual cues.

With respect to the interplay of the acoustic cues, Francis *et al.* (2000) studied simultaneous variation and cue-weighting in the stop voicing distinction in English. They showed that if the stop closure duration in the word *rabid* is increased to more than 70 ms, then listeners hear the word *rapid*, but *only* when there is no voicing, i.e., in the absence of voicing maintenance during stop closure. For English stop consonant place and manner identification, Bailey and Summerfield (1980) showed that in fricative-stop syllable initial clusters (e.g., *star*, *spar*, or *scar*) all of the acoustic differences measured in natural productions are able to provide correct stop perception results. Remarkably, no single cue was necessary and many different cues in combination were sufficient for a correct perception of place and manner of stops. Evidence suggests that multiple processes are responsible for both generating and maintaining accurate perception in the face of all the variability encountered in speech. Furthermore, there is some evidence that some listeners give more weight to specific perceptual cues than other listeners. Redundant cues often enter into trading relationships that increase the magnitude of certain cues while decreasing the magnitude of others, i.e., the listener decision is the result of a certain cue-trading of the perceptual system. The complexity increases when taking into account language differences: Oglesbee (2008) showed that, in a multidimensional stop

[a)]Author to whom correspondence should be addressed. Electronic mail: danielpape@ua.pt

 0001-4966/2014/136(3)/1334/10/$30.00

categorization task, a comparison of listeners from different languages results in different preferences for the important perceptual cues, thus pointing to a language-dependency for both selection and weighting of available stop contrast cues.

Even for various languages of the same family like Italian, Spanish, and European Portuguese (EP), the role of acoustic cues and cue-trading is rather difficult to predict. For example, Italian shows a stable *voicing maintenance* (the presence of voicing throughout the complete stop closure) for all of their word-medial phonologically voiced stops, as shown by Shih *et al.* (1999) in a cross-linguistic database comparison, and recently by Pape and Jesus (2014b). EP, in contrast, has completely different *voicing maintenance* patterns for their phonologically voiced stops. Recent findings (Lousada *et al.*, 2010; Pape and Jesus, 2014a), showed that EP phonologically voiced stops often have no discernible burst, and, more importantly, most of the EP phonologically voiced stops are produced as devoiced (Pape and Jesus, 2014a). Even in the matched Italian and EP data in Pape and Jesus (2014b) where we controlled for all relevant acoustic and linguistic parameters, we could show that devoicing in EP is significantly higher than in Italian. Thus, it would be reasonable to assume that the absence of voicing maintenance during stop closure and the missing burst would lead EP listeners to identify the phonologically voiced consonants as voiceless. However, it is probable that other cues take over to guarantee a robust perceptual outcome. Results from a previous perceptual study (Veloso, 1995) on contrasting voiced and voiceless EP stops showed that when replacing varying stop durations with white noise in isolated VCV (vowel-consonant-vowel) sequences, EP native listeners' responses were distinct for different stop durations. It has to be noted that the results however were limited to an increase of listeners' voiceless classifications for longer stop durations. Furthermore, none of the other possible acoustic cues were removed or controlled for in this study, so Veloso (1995) only tested the variation in stop duration and, obviously, the effects of the complete absence of the voicing maintenance cue. Thus, from Veloso (1995) it is very difficult to extract valid conclusions about the use of multiple acoustic cues for voicing distinction in EP. Speech production results for EP, on the other hand, point to the hypothesis that stop voicing distinctions would not be very different from other Germanic or Romance languages, since stop durations (Martins, 1975; Pape and Jesus, 2014a; Veloso, 1995; Viana, 1984) and preceding vowel durations (Pape and Jesus, 2014a,b) are significantly different when contrasting voiced and voiceless stops. However, the very high percentage of completely devoiced phonologically voiced stops in EP (Lousada *et al.*, 2010; Pape and Jesus, 2014a) is not found in other Romance languages like Italian (Pape and Jesus, 2014b) or Spanish (Shih *et al.*, 1999). In fact, the high overall percentage of devoiced items found in EP production studies is rather comparable to Germanic languages (Shih *et al.*, 1999).

The aim of this study is therefore to disentangle the described controversies for EP by examining the use of several acoustic cues for stop voicing perception, and to shed light on the interplay of the different cues for intervocalic medial stops, all in the absence of a facilitating burst (the presence of a burst allows for a VOT and thus a robust cue to voicing). In other words, we are interested in finding out how the perceptual system evaluates voiced/voiceless distinction, which of the available cues is chosen, and how a possible interplay of several available cues is ordered. For this purpose, we will use voicing patterns and durational values extracted from real EP speech productions (Pape and Jesus, 2014a; Pape *et al.*, 2012), with the goal of examining the actual use and interaction (cue-weighting) of the perceptual cues vowel duration, stop duration, and voicing maintenance.

II. METHOD

A. Listeners

We recruited 38 female native EP listeners. The mean age was 20 yrs (standard deviation 1.5 yrs). All listeners were in the first or second year of their university education (Health Sciences) at the University of Aveiro (Portugal). They did not receive course credit or financial compensation for their participation. None of the listeners reported speech or hearing problems. The majority of the listeners ($n = 23$) came from the region of Aveiro/Porto, 7 listeners came from the neighboring Viseu/Guarda area, 4 listeners from the Lisbon region, 3 listeners from the Azores, and 1 listener came from the Madeira Islands.

B. Stimulus generation

We used biomechanical modeling to generate the perceptual stimuli for the experiments. The reason for using biomechanical modeling in contrast to, for example, formant synthesizers, lies in the ability of the biomechanical models to generate physically realistic trajectories between consecutive phonemes. In other words, articulatory trajectories are not linearly interpolated, as is normally the case with other synthesis approaches. Research on trajectories has shown that the characteristics of curved paths are explained by anatomical factors and muscle mechanics, for arm movements (Flanagan *et al.*, 1993; Gribble and Ostry, 1998, 2000) as well as for speech movements (Iskarous *et al.*, 2010; Perrier and Fuchs, 2008; Perrier *et al.*, 2003; Tasko and Westbury, 2002). So, biomechanical modeling, in contrast to other synthesis approaches, has the advantage that all obtained tongue movements, trajectories, and phoneme targets are comparable to natural speech. This allows the manipulation of glottal source parameters while maintaining articulatory realism. Thus, the use of biomechanical modeling is the best compromise to guarantee highly realistic perceptual stimuli for our experiments, without the risk of missing hidden perceptual cues (which cannot be controlled for) when using, for example, manipulated natural speech.

The biomechanical model described in Perrier *et al.* (2003) allows for the independent parameterization of vowel and consonant durations (implemented via holding and transition times). The correctness of the model's vowel and consonant targets has been verified by EMMA (Electromagnetic Midsagittal Articulography) recordings (Perrier *et al.*, 2003). Following the generation of appropriate

muscle commands for the targets and the holding and transition times, the two-dimensional (mid-sagittal) tongue contours were transformed into area functions (Perrier *et al*., 1992) for each timeframe. Each area function was then acoustically synthesized by a Kelly-Lochbaum (Kelly and Lochbaum, 1962) model developed by Story (2004, 2005) and excited by a three-mass vocal fold model (Titze, 1984). All stimuli were synthesized with flat *F*0 contours, and the voicing extinction parts of the signal were implemented with damping functions of the acoustical signal. We did not model the stop burst noise (i.e., stop release noises are completely missing). Further details on the biomechanical modeling and generation of the EP stimuli are given in Pape *et al*. (2012). Each durational difference between the various stimuli (both vowel and stop) was synthesized using a modification of the original biomechanical model, using the model's transition and holding commands for manipulation. Thus, no signal processing of the resulting acoustical model was necessary. All durations of stimuli were modeled by adequate durational differences, corresponding to physically correct human speech movements.

In sum, the speech material generated for the perceptual experiments consists of biomechanically modeled stimuli acoustically synthesized with a parametric model of the vocal tract and a three-mass vocal fold model. The biomechanical modeling has the main advantage that all tongue movements, trajectories, and phoneme targets are comparable to natural speech, with the additional possibility to manipulate relevant temporal and glottal source parameters while at the same time maintaining articulatory realism. This guarantees highly realistic perceptual stimuli where we can independently control parameters such as duration, transition, and target.

C. Experimental design

We aimed to examine three different factors known to influence the perception of stop voicing: Stop duration, preceding vowel duration, and voicing maintenance during stop closure. Each factor was laid out in a continuum with several levels and was combined with all levels of the other factors (i.e., fully crossed and non-adaptive design). The values of the extreme points of the continua of the three factors were determined according to the values of an extensive EP production database of six speakers (ten repetitions of all EP stops, three vowel contexts) of the Aveiro/Porto region (see Pape and Jesus, 2014a). Table I shows the means and standard deviations for both the preceding vowel and stop durations of the EP velar stops.

TABLE I. Mean durations (standard deviation in parentheses) of the velar stop /k/ /g/ in the EP production database (Pape and Jesus, 2014a). Shown are the averaged values of six speakers for the medial velar stop /CVCV/ in the frame sentence "Diga CVCV outra vez" for the /a/ and /o/ context.

	Vowel context	Preceding vowel duration [ms]	Stop duration [ms]
/g/	/a/	132 (35)	95 (17)
	/o/	130 (20)	104 (20)
/k/	/a/	74 (34)	149 (18)
	/o/	70 (20)	157 (18)

The three factors and their level values were determined as follows (see Table II):

(1) Stop duration: Mean durations (rounded to the closest decimal) of the unvoiced and phonologically voiced velar stops /k ɡ/ in the vowel contexts /a o/ were taken as the range limits of the stop duration continuum, i.e., 100 ms (mean of the voiced stop) and 150 ms (mean of the unvoiced stop). One intermediate value (125 ms) was introduced.
(2) Vowel duration: Mean durations (rounded to the closest decimal) of the preceding vowels /a o/ of the unvoiced/voiced velar stops /k ɡ/ were taken as the range limits of the vowel duration continuum, i.e., 70 ms (mean of the preceding vowel of the unvoiced velar stop) and 130 ms (mean of the preceding vowel of the voiced velar stop). One intermediate value (100 ms) was introduced. Table I shows the duration values from our EP production study (Pape and Jesus, 2014a) for the vowel context /a o/ (the vowel durations entered into the parameter set of the biomechanical model were the durations of /a/ when the resulting stimuli were /aCa/ or the durations of /o/ when the resulting stimuli were /oCo/).
(3) Voicing: The voicing maintenance continuum was defined by the two endpoints *fully voiced* and *fully devoiced/unvoiced*. For the intermediate values, five conditions were defined (12.5%, 25%, 37.5%, 50%, and 75%) at which the stop voicing ceases and thus the stop devoicing begins (and remains until its offset). The unequal step sizes result from the hypothesis that the perceptual differences would be smaller toward lower voicing percentages of the stimuli, so smaller step sizes for higher voicing percentages were excluded to obtain a reduced total number of stimuli. Figure 1 shows the waveforms for the *fully devoiced/unvoiced* condition (0%) in the bottom panel, the *fully voiced* condition (100%) in the top panel, and intermediate voicing maintenance conditions in the middle (37.5% and 75%). Please note that the *fully devoiced* condition denotes different underlying control mechanisms than the *unvoiced* condition (Jesus and Shadle, 2002, 2003), although the result, i.e., the voicing maintenance, is identical in both conditions.

We thus obtained a three-factor design with $3 \times 3 \times 7$ levels of the corresponding continua. These were embedded

TABLE II. The three factors (and their step size) used to construct the three continua in the design of the perceptual stimuli. Stop duration and vowel duration are given in milliseconds, the voicing maintenance is given as the percentage of the voiced part in reference to complete stop duration.

Steps	Vowel duration [ms]	Stop duration [ms]	Stop voicing [%]
1	70	100	0 (fully devoiced/unvoiced)
2	100	125	12.5
3	130	150	25
4			37.5
5			50
6			75
7			100 (fully voiced)

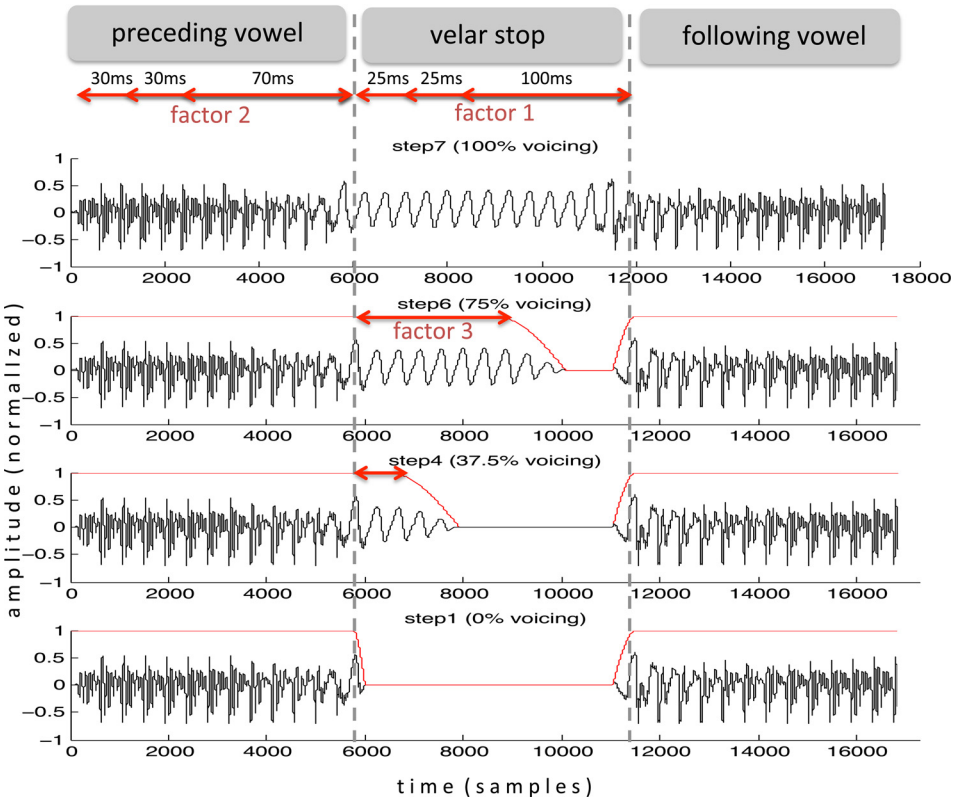

FIG. 1. (Color online) Waveforms for the synthesized VCV stimuli for four steps of the stop *voicing maintenance* continuum: The top and bottom panels show the two extremes (top = fully voiced; bottom = fully devoiced/unvoiced); the two intermediate panels show the partly voiced conditions (second panel step6 = 75% voiced; third panel step4 = 37.5% voiced). The arrows display the three factors used in the perception experiment: Stop duration (factor 1), vowel duration (factor 2), and stop voicing maintenance (factor 3).

in two different vowel contexts /a o/. Table II shows the values and step sizes of the three factors selected for the perceptual experiments.

D. Experimental setup and listeners' tasks

We designed two perception experiments, one as a speeded discrimination task and one as an identification task. The two experiments served different aims and approaches to the expected perceptual outcomes: The discrimination task was designed to test for the listeners' *ability* to distinguish between a set of several *voicing maintenance* differences during stop closure (hence to test if the stimuli are perceptually discriminable). In contrast, the identification experiment was designed to allow the integration of several acoustic cues to obtain a given voiced/voiceless decision. Thus, in the identification experiment we test whether the listener relies on the use of *voicing maintenance* cues to obtain a given perceptual construct. Hence, the main difference between the discrimination and the identification task is whether the perceptual system of the listener *is able* to perceive a given voicing maintenance differences (discrimination) versus whether the listener *relies* on this cue in his/her voiced/voiceless distinction (identification). We carried out the two experiments one after the other, divided by a short break. This resulted in two separate experiments carried out in sequence.

In the *discrimination task* we are interested in the discrimination threshold of *voicing maintenance* during stop closure. For this reason we selected for the perceptual stimuli the vowel and stop closure durations that would correspond to ambiguous listening conditions, i.e., situations where the listener is not able to extract additional perceptual cues from vowel or stop duration and thus must generate a perceptual output based solely on other acoustic cues. We therefore selected durational values for the VCV stimuli that are intermediate between the voiced and voiceless velar stop based on the data derived from our EP production database (i.e., vowel duration 100 ms and stop closure duration 125 ms). We tested all possible pairs of *voicing maintenance* differences (0% with 13%, 0% with 25%,…, 75% with 100%) with the two different context vowels /a o/ (thus providing the listener with stops both in the /a/ and in the /o/ context). We also included pairs of *identical stimuli* (e.g., testing 0% with 0% *voicing maintenance*, or 50% with 50%). We ran three repetitions of the complete stimuli set in randomized order. In total, the complete discrimination experiment consisted of 144 trial pairs to be judged by the listener [21 possible combinations of pairs of different *voicing maintenances* plus 3 pairs of *identical stimuli* multiplied by two vowel conditions (/**a**C**a**/ and /**o**C**o**/) multiplied by 3 repetitions]. The listeners' task was to decide whether the second stimulus was identical to the first stimulus (AX task). The time interval between A and X was 100 ms to guarantee a low memory load (Pisoni and Trash, 1974). The listeners were informed that they would hear synthetic VCV items. There was a practice session of 25 stimuli prior to the beginning of the experiment. We emphasized speed of response, asking listeners to respond as quickly as possible, but also as accurately as possible. The average time (over all listeners) to perform the discrimination experiment was 10 min. Trial repetition was not possible.

The *identification task* focuses on differences in voiced/unvoiced perception based on *voicing maintenance*, stop

closure duration, and contextual *vowel duration* differences. This identification task integrates the three available cues for stop voicing perception and thus taps into the higher perceptual streams. In this experiment we tested all possible combinations of contextual vowel duration (70 ms, 100 ms, and 130 ms) with all combinations of stop duration (100 ms, 125 ms, and 150 ms) and all voicing maintenance steps (0%, 12.5%, 25%, 37.5%, 50%, 75%, and 100%). The experiment was performed in two different vowel conditions (/a/ and /o/). We ran five repetitions of the complete stimuli set for each of the listeners in randomized order. In sum, we obtained 126 stimuli repeated 5 times for a total of 630 trials (3 *vowel durations* multiplied by 3 *closure durations* multiplied by 7 *voicing maintenance* conditions multiplied by 2 *vowel* identities and multiplied by 5 *repetitions*). The average time (over all listeners) to perform the task was 20 min. There was a practice session of 25 stimuli prior to the beginning of the main experiment. Listeners were informed that they would hear synthetic VCV items, and their task was to identify whether the consonant was /g/ or /k/ (forced choice). We emphasized speed of response, asking listeners to respond as quickly as possible, but also as accurately as possible. Trial repetition was not possible.

For the experimental setup, we used open headphones with a linear frequency response (Sennheiser HD 600, Wedemark Wennebostel, Germany) connected to the internal headphone output of a notebook computer (no other processes running, all networking interfaces disabled). Listeners' responses were collected by means of mouse clicks placed on different buttons on the screen. Listeners were seated in a soundproof room (size around 3 m × 2 m) at University of Aveiro's Speech, Language and Hearing Laboratory (SLHlab). We used *Alvin* v 1.27 (Hillenbrand and Gayvert, 2005) open source software for stimulus and visual presentation. The computer screen for the *identification task* showed two buttons (labeled "g" and "k"), one on the left side and one on the right side (at identical distances), around a "next" button at the screen center. After selecting their response choice (i.e., "g" or "k"), the listeners had to click on the "next" button to proceed to the next stimulus, thus placing the cursor at the exact center of the screen before the next stimulus presentation (guaranteeing identical distances for the two answer possibilities). The *discrimination task* had an identical setup, differing only in terms of the two choice buttons (labeled "same" and "different"). All button options and accompanying text were written in Portuguese in order to not confuse listeners' internal language representation. The placement of all buttons was rotated 180° for one-half of the listener population, thus counterbalancing biases of horizontal movement and listener preference.

E. Statistical analysis

To test for the statistical validity of the listeners' response patterns in our discrimination and identification experiments we performed a series of Logit models with mixed effects [Generalized Linear Mixed Models (GLMM), function lmer (Bates *et al.*, 2011)] in the R environment (RDCT, 2010). Logit models are suited for dependent variables with binomial distributions (using *z*-scores). This allowed statistical modeling based on binary decisions (Baayen, 2008), as is the case with our listener responses (for the identification experiment: /g/ or /k/ response; for the discrimination experiment: same or different response).

III. RESULTS

When analyzing the responsiveness of all 38 listeners to the 3 examined acoustic cues (*vowel duration*, *stop duration*, and *voicing maintenance*), 31 listeners showed a response pattern that differed from completely random response in the identification experiment. A listener was regarded as responsive and thus included in the analyses when the individual response patterns showed substantial deviation from a 50% random response pattern for at least one of the three acoustic cues. There was no difference in the response patterns (both identification and discrimination) when comparing listeners from Mainland Portugal with those from the Azores and Madeira Islands. We excluded the 7 listeners with random response patterns (18.4% of the recruited population) from both the discrimination and identification experiments and report in the following the results for the 31 responsive listeners. Please note that the percentage of random response listeners is higher than normally expected (nearly 20%), which is due to the rather difficult identification task for the listeners (e.g., absence of burst).

A. Discrimination: Voicing maintenance differences for intermediate durational values

We computed a Logit model (GLMM) based on the binomial listener response (same or different) as the dependent variable. Fixed factors were *voicing maintenance difference* between stimulus A and stimulus X [0% = identical; 13%; 25%; 38%; 50%; 63%; 75%; 88%; 100%], *vowel identity* [a; o], and *repetition* number [1; 2; 3]. The factors *voicing maintenance difference* and *repetition* were centered and scaled (*z*-transformed). Listener number was regarded as a random factor and was incorporated with random intercept and random slope into the GLMM. We did not use a fully randomized model (i.e., we did not estimate the effect of each of the predictors on individual-level slopes). We found that the *null* model (consisting of the factors *vowel identity*, *repetition*, and the random factor) was highly significantly different ($chi^2 = 482$; $p < 0.001$) from the *full* model (consisting of all *null* model factors plus the factor *voicing maintenance difference*). For the full model, only the factor *voicing maintenance difference* was highly significant ($z = 20.9$, $p < 0.001$), but not the factors *vowel identity* ($z = 1.4$, $p < 0.17$) and *repetition* ($z = -0.7$, $p < 0.49$). In sum, differences in the *voicing maintenance* during stop closure between the A and the X stimuli have a highly significant effect on the listeners' same/different responses, showing that listeners are able to discriminate between different levels of *voicing maintenance* across stimuli pairs. All seven levels of the *voicing maintenance* difference were significantly different from all other levels (Bonferroni corrected), showing that listeners are highly efficient in

differentiating among all seven levels of *voicing maintenance differences*.

Since we included *change trials* (i.e., a physical difference between stimuli within a pair) and *no-change trials* (i.e., no difference between stimuli—"identical" stimuli), we could—according to signal detection theory—calculate the A-prime (A′) scores based on the proportion of *hits* (listener perceives a difference when presented with a change trial) and *false alarms* (listener perceives a difference when presented with an identical stimuli) (Snodgrass and Hayden, 1985). An A′ score of 1.0 indicates perfect discrimination (correct responses to all change and no-change trials), whereas as a score of 0.5 or lower indicates a lack of perceptual phonetic sensitivity. In other words, the value of A′ increases with the number of correct responses to *change trial* stimuli *and* correct rejections of the identical stimuli. Figure 2 shows the A′ values as a function of the *voicing difference* between the presented stimulus A and stimulus X. A voicing difference of 100% on the *x* axis is the representation of a pair of *fully voiced (A) and fully unvoiced (X)* stimuli (or vice versa); a difference of 25% can be the presentation of a pair (1) of a 75% voiced stimulus (A) with a fully voiced stimulus (X) or (2) of a fully devoiced stimulus (A) with a 25% voiced stimulus (X). As can be seen in Fig. 2, discrimination scores increase roughly monotonically with an increase in voicing maintenance difference.

However, the 50% uncertainty point across all listeners, i.e., the turning point where stimuli pairs are equally perceived as being identical or different, is reached only for higher voicing differences. Thus, over all listeners, the voicing maintenance difference between the A stimulus and the X stimulus has to exceed 75% to be reliably perceived as phonetically different. In other words, a stimulus with *13% voicing* during closure is clearly perceived as different from a *fully voiced* stimulus (*100% voicing*) by the majority of the listeners, but smaller differences, as for example *25% voicing* compared to *75% voicing*, are not reliably perceived as different.

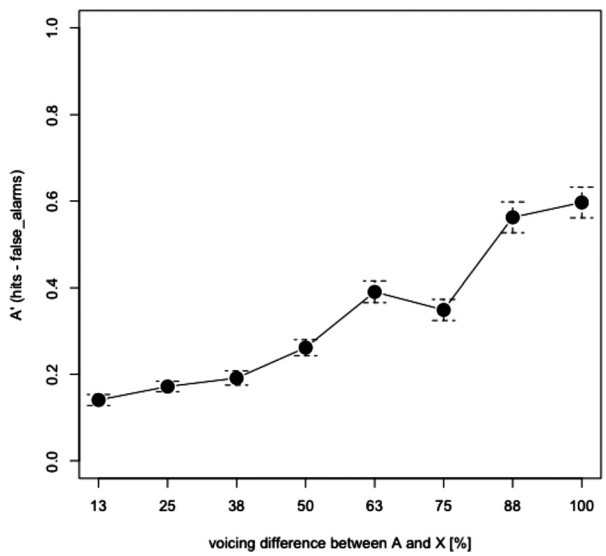

FIG. 2. Discrimination of voicing maintenance differences in the stimuli pair (*x* axis) for intermediate durational values between /g/ and /k/. Shown are means and standard errors of the value A′ (A-prime, see text for explanation) relative to the voicing difference of each AX stimulus pair.

B. Identification: Effects and interplay of the acoustic cues

Figure 3 shows a first overview the effect of the three factors (*vowel duration* in different panels, *stop duration* as different lines in each panel, *voicing maintenance* on the *x* axis) on the mean percentage of listeners' /g/ responses (over all 31 responsive listeners). For the factor *voicing maintenance* two effects can be observed: First, there is a clear bias for listeners to prefer /g/ responses, with a lack of complete /k/ responses (i.e., a 0%–20% *voiced* response probability) and a lack of 100% /g/ responses. This lack of robust /g/ and /k/ responses again shows that the perception task is rather difficult for the listeners. Second, increasing *voicing maintenance* leads to an increase of voiced responses for all *vowel duration* and *stop duration* conditions. Furthermore, increasing *stop duration* leads to a decrease in the probability of voiced responses, but mainly for low *voicing maintenance* values (i.e., for highly devoiced or unvoiced items). Increasing *vowel duration* (panels from left to right) results in an increase in the /g/ response probability, except for high *voicing maintenance* values. Increasing the *voicing maintenance* thus increases the /g/ response probability for all *vowel durations* and *stop durations*, however with different magnitudes. *Stop duration* and *vowel duration* have a strong influence on voicing decisions, but this effect is limited to low *voicing maintenance* values. For higher *voicing maintenance* percentages (over 75%), we observed a ceiling effect for all *stop durations* and *vowel durations*. It seems that the *voicing maintenance* cue is very strong here and overrules the other two acoustic cues.

To test the statistical validity of these observations, we computed a Logit model (GLMM) for the listeners' binomial response as the dependent variable (i.e., /g/ or /k/ response). We tested with a $p < 0.05$ significance threshold the effects of the fixed factors *vowel duration*, [70; 100; 130] *stop duration*, [100; 125; 150] *voicing maintenance* percentage during stop closure, [0; 13; 25; 38; 50; 75; 100] *vowel identity* [a; o], and *repetition* [1; 2; 3; 4; 5]. *Listener* was selected as a random factor and was included with random intercepts and random slopes for all factors. We included all possible (two- and three-way) interactions. Since we are only interested in the effects of *voicing maintenance*, *vowel duration*, and *stop duration*, we consequently tested the *null* model (including only the other fixed factors and the random factor) against the *full* model (including the fixed factors of the *null* model and additionally the factors *voicing maintenance*, *vowel duration*, and *closure duration* with all interactions). The full model was significantly different from the null model ($chi^2 = 4191$, $p < 0.001$), thus showing that our main three fixed effects significantly contributed to the model. In the full model, the factors *voicing maintenance* ($z = 10.4$, $p < 0.001$), *vowel duration* ($z = 9.8$, $p < 0.001$) and *vowel identity* ($z = -3.8$, $p < 0.001$) were significant, but not *stop duration* ($z = -1.4$, $p = 0.16$) or *repetition* ($z = -1.3$, $p = 0.19$). With respect to vowel identity, the low context vowel /a/ showed higher /g/ responses than the /o/ context vowel. There were no (two- or three-way) interactions between the significant factors. To illustrate the goodness of

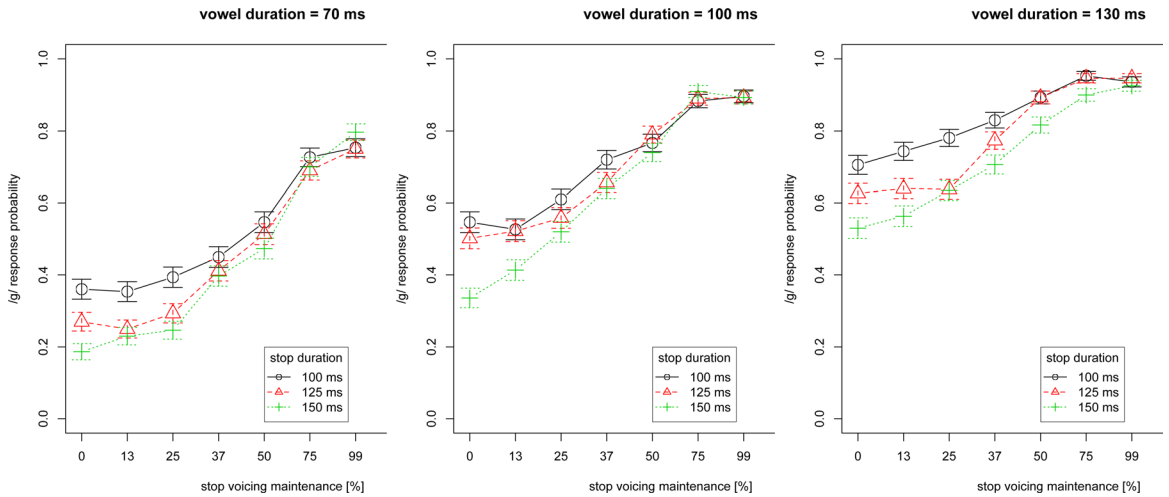

FIG. 3. (Color online) Percentages of /g/ responses (y axis) over all responsive listeners ($n = 31$) with respect to the factor *voicing maintenance* (in percent of the stop closure duration, x axis). The three panels split the data by vowel duration, and within each panel the different lines correspond to the three stop durations.

fit of the selected full GLMM, Fig. 4 plots (in separate panels for the three significant *vowel durations*) the means of the raw perceptual data (dots) against the fit of the full (linear mixed) model (dashed lines). It can be seen that the chosen GLMM predicts very well the listener responses observed in the raw perceptual data.

Although the factor *stop duration* was not significant in statistical analysis, Fig. 3 suggests that for each *vowel duration* there is an influence of stop duration on the perceptual responses. This could be due to the fact that not the absolute durations, but rather the relative ones play a role. This is consistent with the idea that relative timing matters in differentiating VCV sequences (Tuller and Kelso, 1984). We attempted to explore the complex interaction of *voicing maintenance*, *vowel duration*, and *stop duration* from a different perspective by computing the responses to the *voicing maintenance* cue as a function of both *vowel duration* and *stop duration*. For this reason, we defined a ratio r as the proportion between *vowel duration* and *stop duration* ($r = \mathrm{VL/CL}$). This factor r is based on the observation that in

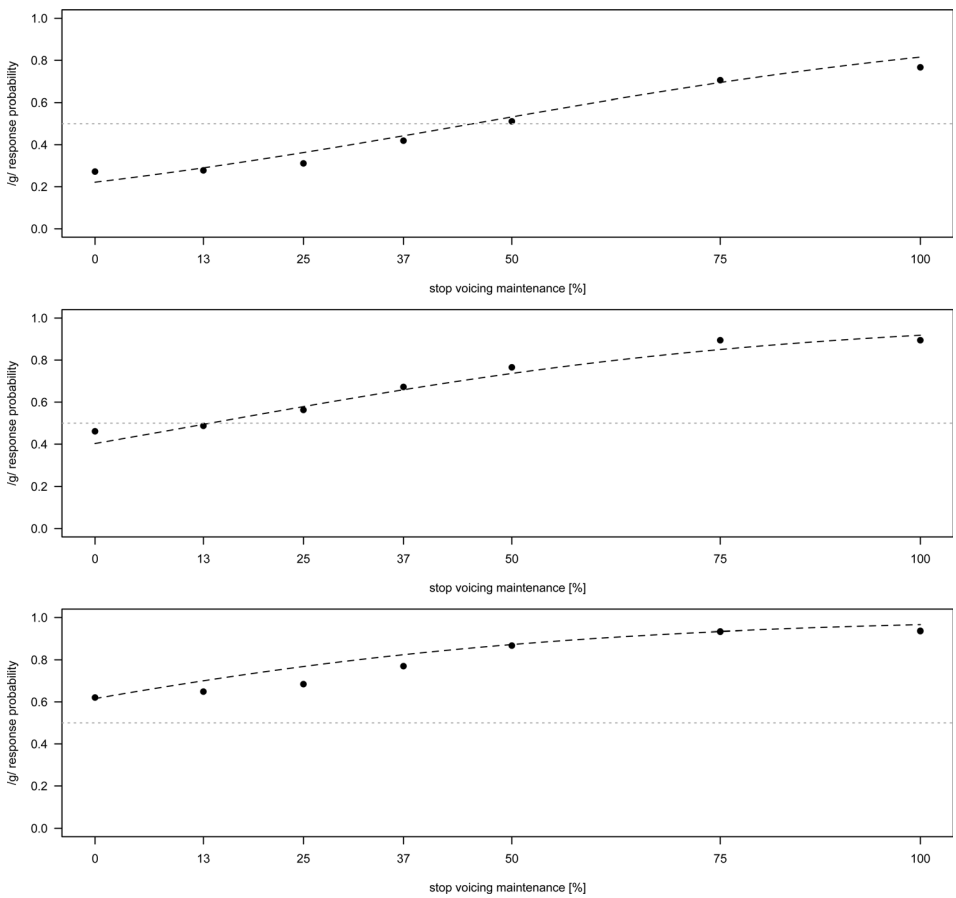

FIG. 4. Fit of the full Linear Mixed Model (dashed lines) against the means over all listeners for each of the seven *voicing maintenance* levels (x axis). The y axis shows the probability of voiced /g/ responses. The different panels split the data by the three *vowel durations:* Top panel 70 ms, middle panel 100 ms, and bottom panel 130 ms *vowel duration*. The 0.5 probability is shown as a gray dotted line.

speech production data the sum of preceding vowel duration and stop duration is similar for voiced stops and voiceless stops (see Table I: Around 220 ms for both the voiced and the voiceless velar stop). For speech perception, both from the literature and from the results presented in Fig. 3, it is evident that increasing *vowel duration* and/or decreasing *stop duration* leads to higher probabilities of a voiced listener response. Thus, high values of the factor *r* can be associated with longer relative vowel durations and/or shorter relative stop durations (thus durational values linked with prototypical /g/ values), while small *r* values are characterized by shorter relative vowel durations and/or increased relative stop durations (thus prototypical /k/ values). Figure 5 plots the *voicing maintenance* data as a function of factor *r*. It can be seen that increasing *r*—or moving toward the durations of a prototypical /g/—induces a noticeable increase in the voiced identifications except for the fully voiced stimuli set. This could be due to the fact that for the fully voiced items we observe a ceiling effect toward listeners' voiced response, independent of the stimulus *vowel duration* and *stop duration*. In contrast, for fully devoiced/unvoiced items there is a strong influence of the relative duration, systematically reaching a probability higher than 0.5 at a value of $r = 0.8$. Further, when comparing the left and right panel of Fig. 5 it can be seen that the influence of the contextual vowel identity on the listener responses, although significant in the statistical analysis, can be neglected for all *r* ratios and all three voicing maintenance conditions.

IV. DISCUSSION AND CONCLUSION

The experiments in this paper showed that for the perception of intervocalic velar stops EP listeners are sensitive to the acoustic cue *voicing maintenance*. In the discrimination experiment we showed that *voicing maintenance* is a discriminating factor if differences exceed a certain threshold (see the discrimination scores in Fig. 2 for low voicing). Furthermore, the second perception test provided evidence for the role of the cue *voicing maintenance* in voiced versus unvoiced identification. In this regard, the perceptual results for EP are comparable to those obtained for other languages, for example English (Lisker, 1986).

This result contradicts the observation that in EP the phonetic realizations of phonologically voiced stops are often highly devoiced throughout the complete duration of their stop closure (Pape and Jesus, 2014a). From these production results, one could assume that for EP stop voicing perception *vowel duration* and/or *stop duration* are more important acoustic cues than *voicing maintenance*, or that all these factors interact in a complex manner that does not give *voicing maintenance* the strongest weight in the identification process. This assumption could be supported by the data presentation in Fig. 5 that shows the influence of the vowel duration measured in relation to the stop duration (ratio *r*): For completely devoiced items (0% voicing maintenance line) a strong influence of *r* (and thus of vowel duration and stop duration) on the listeners' responses can be observed. However, this influence becomes negligible for fully voiced items (100% voicing maintenance). Thus, voicing maintenance is the dominant cue if the presented stimulus is fully voiced, and is thus the major cue triggering the voicing decision of the EP listeners. But if the presented stimuli are (fully) devoiced, the two other acoustic cues (*vowel duration*, *stop duration*) take over by triggering the voiced/voiceless listener responses based on the extracted phoneme durations. In other words, we encounter cue-weighting among different acoustic cues in the perception of EP stop voicing. In sum, these observations support the hypothesis that voicing maintenance is a major but not a required cue for stop voicing perception in EP.

What is striking about these interactions is the observation that our statistical analysis showed that only the two acoustic cues *voicing maintenance* and contextual *vowel duration*, but not *stop duration*, have a significant effect on voicing distinction. The absence of statistical significance of *stop duration* is in contrast to previous findings in the perceptual study of Veloso (1995). However, in Veloso (1995) only the *stop duration* variation was tested, without controlling for other acoustic cues. In addition, there was no voicing during closure (i.e., the complete stop duration was masked

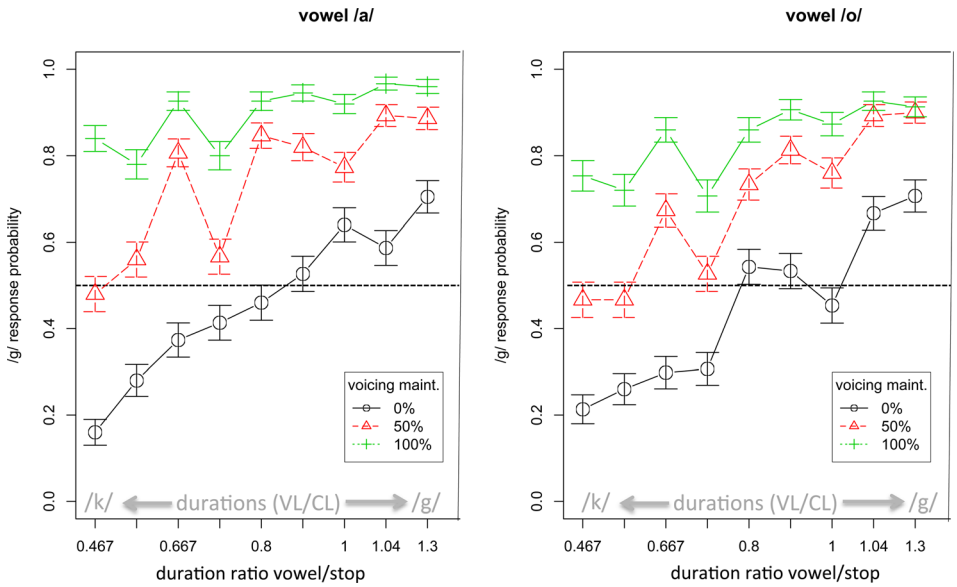

FIG. 5. (Color online) *Voicing maintenance* cue (separate lines) as a function of factor *r* (*x* axis, see text for explanation). Higher *r* values are associated with prototypical /g/ vowel and stop durations, low *r* values correspond to prototypical /k/ durations. Mean and standard errors for all **/aCa/** stimuli are shown at the left panel and the **/oCo/** stimuli are shown at the right panel. The *y* axis presents the mean probability (over all listeners) that a voiced stop /g/ was perceived.

by white noise and thus only the duration was varied in length), so the results of Veloso (1995) are difficult to explain. Since the perceptual bias was limited to the listeners' "voiceless" responses (i.e., no perceptual shift toward "voiced" responses was found when varying the stop duration), the results of Veloso (1995) could be due to (1) the missing voicing during stop closure, (2) the very short stop closures, or (3) the lack of control of other cues, such as preceding vowel duration or VOT.

The absence of an effect of *stop duration* on stop voicing distinction also contrasts with findings for other languages [e.g., Francis *et al*. (2000) for English]. However, our findings that EP listeners are highly sensitive to the *voicing maintenance* and contextual *vowel duration* cues are in line with previous findings for other languages [see, e.g., Lisker (1986) for voicing maintenance and Raphael (1972) for vowel duration].

Another interesting finding is the bias of all EP listeners toward voiced responses, as can be seen in Fig. 3 even for the fully devoiced/unvoiced condition and durational values prototypical of a voiceless stop. For example, over all listeners, the response probability is 0.2 that a voiced stop will be perceived when listeners are presented with 0% *voicing maintenance* and the prototypical /k/ *vowel duration* (70 ms) and *stop duration* (150 ms). This response probability is the lowest value we could find in our data. In other words, over all listeners we never obtained a stable /k/ response floor effect (0% response probability). We assume that this listener bias is due to the missing burst in the presented stimuli, and thus the problem of extracting a stable VOT cue. However, as described in Sec. I, a large amount of EP data for voiced stops, and even for voiceless stops, does not show a discernible burst (Lousada *et al*., 2010; Pape and Jesus, 2014a), so it is not clear how the listeners would rely on a VOT cue in these ambiguous conditions. The stimuli used in our perception experiments were designed to exclude the burst cue (i.e., to exclude both burst and voicing onset time cues) to examine the influence of voicing maintenance and durational cues on stop voicing distinction. Based on other languages, burst and VOT have an important effect on stop voicing distinction (Jessen, 1999; Lasky *et al*., 1975; Lisker, 1986), so clearly the results would have been different if the burst had been included among the perceptual constructs (e.g., it could be the case that the seven non-responsive listeners are only responsive to the burst and/or the VOT cue; however, testing of this hypothesis would require another perceptual experiment).

We conclude that, with the constraint of the missing burst in the EP stimulus set, the results for the identification and the discrimination task show that the *voicing maintenance* cue is strongly used to distinguish voicing, in addition to and in combination with the *vowel duration* cue. In contrast, the *stop duration* cue is not widely used to obtain a robust voicing distinction. *Stop duration* in combination with *vowel duration* (value *r*) seems to play the major role in distinguishing voiced and voiceless stops, but *only* when the stimuli are highly devoiced. Since this devoicing is very often found in the production of EP phonologically voiced stops, it is concluded that in this case *vowel duration* as the major cue and *stop duration* as a minor cue are mainly used for stop voicing distinction. In other words, even with the missing burst and in the complete absence of voicing during stop closure, EP listeners are able to make robust voiced/voiceless decisions based only on *vowel duration* and *stop duration*. It has to be noted though that these robust decisions are limited by ceiling and floor effects we found—we were not able to obtain complete voiced (100% voiced) or unvoiced (100% unvoiced) responses of all EP listeners.

However, if the stimuli to be judged are more or less fully voiced, then the *voicing maintenance* is found to be the major cue. In this case it overrides the *vowel duration* and *stop duration* cues, and guarantees, even in the absence of a facilitating burst, a stable voiced response of all EP listeners. In sum, this study constitutes new evidence that, in absence of a facilitating burst, multiple acoustic cues are used and combined with different cue-weighting to obtain a stable stop voicing distinction.

ACKNOWLEDGMENTS

This work was partially funded by National Funds through FCT—Foundation for Science and Technology—in the context of the project PEst-OE/EEI/UI0127/2014 to IEETA, and the postdoctoral fellowship from FCT (Portugal) SFRH/BPD/48002/2008 to D.P. The authors also thank all the subjects for their participation in the experiments. We thank Pascal Perrier for valuable comments on biomechanical modeling and the manuscript. We thank Roger Mundry for his help with the statistical analysis. We would like to thank Benjamin Munson for his insightful critiques.

Baayen, H. R. (**2008**). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R* (Cambridge University Press, Cambridge), 368 p.

Bailey, P. J., and Summerfield, Q. (**1980**). "Information in speech: Observations on the perception of [s]-stop clusters," J. Exp. Psychol. **6**, 536–563.

Bates, D., Maechler, M., and Bolker, B. (**2011**). "lme4: Linear mixed-effects models using S4 classes," R package version 0.999375-42, http://CRAN.R-project.org/package=lme4 (Last viewed August 10, 2013).

Cuartero, N. (**2002**). "Voicing assimilation in Catalan and English," Ph.D. thesis, Universitat Autonoma de Barcelona, Barcelona, Spain.

Flanagan, J. R., Ostry, D. J., and Feldman, A. G. (**1993**). "Control of trajectory modifications in target-directed reaching," J. Motor Behav. **25**, 140–152.

Francis, A., Baldwin, K., and Nusbaum, C. (**2000**). "Effects of training on attention to acoustic cues," Atten., Percept., Psychophys. **62**, 1668–1680.

Gribble, P. L., and Ostry, D. J. (**1998**). "Are complex control signals required for human arm movements?," J. Neurophysiol. **79**(3), 1409–1424.

Gribble, P. L., and Ostry, D. J. (**2000**). "Compensation for loads during arm movements using equilibrium-point control," Exp. Brain Res. **135**, 474–482.

Hillenbrand, J. M., and Gayvert, R. T. (**2005**). "Open source software for experiment design and control," J. Speech, Lang., Hear. Res. **48**(1), 45–60.

Iskarous, K., Nam, H., and Whalen, D. H. (**2010**). "Perception of articulatory dynamics from acoustic signatures," J. Acoust. Soc. Am. **127**(6), 3717–3728.

Jessen, M. (**1999**). *Phonetics and Phonology of Tense and Lax Obstruents* (John Benjamins Publishing, Amsterdam), 394 p. (in German).

Jesus, L., and Shadle, C. (**2002**). "A parametric study of the spectral characteristics of fricatives," J. Phon. **30**, 437–464.

Jesus, L., and Shadle, C. (**2003**). "Devoicing measures of EP fricatives," in *Computational Processing of the Portuguese Language*, edited by N. Mamed, J. Baptista, I. Trancoso, and M. Nunes (Springer, Berlin), pp. 1–8.
Kelly, J., and Lochbaum, C. (**1962**). "Speech synthesis," in *Proceedings of the Fourth International Congress on Acoustics*, Copenhagen, pp. 1–4.
Lasky, R., Srydal-Lasky, A., and Klien, R. (**1975**). "VOT discrimination by four to six and a half month old infants from Spanish environments," J. Exp. Child Psychol. **20**, 215–225.
Li, F., Mennon, A., and Allen, J. (**2010**). "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," J. Acoust. Soc. Am. **127**(4), 2599–2610.
Lisker, L. (**1986**). " 'Voicing' in English: A Catalogue of Acoustic Features Signaling /b/ Versus /p/ in Trochees," Lang. Speech. **29**(1), 3–11.
Lisker, L., and Abramson, A. S. (**1967**). "The voicing dimension: Some experiments in comparative phonetics," in *Proceedings of the 6th International Congress of Phonetic Sciences (ICPhS 67)*, Prague, pp. 563–567.
Lousada, M., Jesus, L., and Hall, A. (**2010**). "Temporal acoustic correlates of the voicing contrast in European Portuguese stops," J. Int. Phon. Assoc. **40**, 261–275.
Luce, P., and Charles-Luce, J. (**1985**). "Contextual effects on vowel duration, closure duration, and the consonant vowel ratio in speech production," J. Acoust. Soc. Am. **78**, 1949–1957.
Martins, M. (**1975**). "Vogais e consoantes do Português: estatística de ocorrência, duração e intensidade" ("Portuguese vowels and consonants: Frequency statistics, duration and intensity"), Boletim da Filologia, **24**(1–4), 1–11.
Oglesbee, E. N. (**2008**). "Multidimensional stop categorization in English, Spanish, Korean, Japanese, and Canadian French," Ph.D. thesis, Indiana University, Bloomington, IN.
Pape, D., Jesus, L., and Perrier, P. (**2012**). "Constructing physically realistic VCV stimuli for the perception of stop voicing in European Portuguese," in *Computational Processing of the Portuguese Language*, edited by H. Caseli, A. Teixeira, and A. Villavicencio (Springer, Heidelberg), pp. 328–349.
Pape, D., and Jesus, L. (**2014a**). "Stop and fricative devoicing in European Portuguese, Italian and German. Language and Speech," Lang. Speech (in press).
Pape, D., and Jesus, L. M. T. (**2014b**). "Production and perception of velar stop (de)voicing in European Portuguese and Italian," Springer EURASIP J. Audio, Speech, Music Process. **2014**(6), 1–10.
Perrier, P., Boe, L., and Sock, R. (**1992**). "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: Modeling the transition with two sets of coefficients," J. Speech Hear. Res. **114**(3), 53–67.
Perrier, P., and Fuchs, S. (**2008**). "Speed-curvature relations in speech production challenge the one-third power law," J. Neurophysiol. **100**, 1171–1183.
Perrier, P., Payan, Y., Zandipour, M., and Perkell, J. (**2003**). "Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study," J. Acoust. Soc. Am. **114**, 1582–1599.
Pisoni, D., and Trash, J. (**1974**). "Reaction times to comparisons within and across phonetic categories," Percept. Psychophys. **15**, 285–290.
Raphael, L. (**1972**). "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," J. Acoust. Soc. Am. **51**(4:2), 1276–1303.
RDCT (**2010**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing (http://www.R-project.org/). R Development Core Team (Last viewed December 22, 2011).
Shih, C., Möbius, B., and Narasimhan, B. (**1999**). "Contextual effects on consonantal voicing profiles: A cross-linguistic study," in *Proceedings of the 14th International Congress of the Phonetic Sciences (ICPhS 99)*, San Francisco, CA, pp. 989–992.
Snodgrass, J., and Hayden, H. (**1985**). *Human Experimental Psychology* (Oxford University Press, New York), 512 p.
Story, B. (**2004**). "Vowel acoustics for speaking and singing," Acta Acoust. **90**(4), 629–640.
Story, B. (**2005**). "A parametric model of the vocal tract area function for vowel and consonant simulation," J. Acoust. Soc. Am. **117**(5), 3231–3254.
Tasko, S., and Westbury, J. (**2002**). "Defining and measuring speech movement events," J. Speech Hear. Res. **45**, 127–142.
Titze, I. (**1984**). "Parameterization of the glottal area, glottal flow, and vocal fold contact area," J. Acoust. Soc. Am. **75**, 570–580.
Tuller, B., and Kelso, J. (**1984**). "The timing of articulatory gestures. Evidence for relational invariants," J. Acoust. Soc. Am. **76**, 1030–1036.
Veloso, J. (**1995**). "The role of consonantal duration and tenseness in the perception of voicing distinctions of Portuguese stops," in *Proceedings of the 6th International Congress of Phonetic Sciences (ICPhS 95)*, Stockholm, Sweden, pp. 266–269.
Viana, M. (**1984**). "Étude de deux aspects du consonantisme du Portugais: Fricatisation et dévoisement" ("Examining two aspectes of Portuguese consonants: Fricatization and devoicing"), Ph.D. thesis, Université des Sciences Humaines de Strasbourg, Strasbourg, France.