Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 74 (2015) 93–103

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Intervocalic fricative perception in European Portuguese: An articulatory synthesis study

Daniel Pape [a,*], Luis M.T. Jesus [a,b], Peter Birkholz [c]

[a] Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, 3830-193 Aveiro, Portugal
[b] School of Health Sciences (ESSUA), University of Aveiro, 3830-193 Aveiro, Portugal
[c] Institute of Acoustics and Speech Communication, TU Dresden, 01069 Dresden, Germany

Received 4 December 2014; received in revised form 29 August 2015; accepted 1 September 2015
Available online 8 September 2015

Abstract

This study examined the conditions under which European Portuguese (EP) intervocalic fricatives are perceived as voiced/voiceless with respect to varying voiced-to-devoiced fricative portions, durations, pharyngeal widths and glottal heights. Articulatory speech synthesis incorporating a sophisticated noise source model was used to produce /a*C*a/ sequences (C = /f v s z ʃ ʒ/) for two perception experiments (forced choice voiced/voiceless decision). This approach has the main advantage that all relevant articulatory settings can be manipulated independently. The proportion of fricative voicing was varied between *fully devoiced* to *fully voiced*; phoneme durations were varied corresponding to natural EP speech. Glottal height and pharyngeal width each varied amongst three different states. The results showed strong interaction between the cues *phoneme duration* and *voicing maintenance*. A voiced-to-devoiced ratio of only 25% was enough to guarantee robust perception of fricative voicing. *Phoneme duration* and *place of articulation* had a significant effect on listener decisions, but only for voiced-to-devoiced ratios of 25–50%. *Vowel duration*, *pharyngeal width* and *glottis height* had no significant effect. The study provides new evidence for cue-trading in fricative perception. Furthermore, new insights into laryngeal vs. supra-laryngeal gestural coordination are gained which may facilitate the development of fricative models.

Keywords: Fricative synthesis; Fricative voicing; Cue weighting; European Portuguese; Voicing perception

1. Introduction

For speech production, the phonological voicing distinction is defined as the presence or absence of vocal fold vibration during consonant production (Jakobson et al., 1952). For the perception of intervocalic fricative voicing, Stevens et al. (1992) showed in a study of American English that alveolar fricatives, generated with a Klatt (1980) synthesiser, are perceived as voiced if the voiced part of (a partly devoiced) fricative exceeds a duration of 30 ms. Furthermore, when the devoiced part of the (otherwise voiced) fricative exceeds a duration of 60 ms the listener robustly perceives a voiceless fricative. Having established this baseline for the perception of voicing, in their data, Stevens et al. (1992) compared these perception results to an acoustic study conducted beforehand and showed that all of their phonologically voiced fricatives devoiced for less than 60 ms, and 89% of their phonologically voiceless fricatives were voiceless for more than 60 ms. In other words, as expected, the percentage of *voicing maintenance* (defined as the duration of voicing during fricative production in percent of the total fricative duration) was very high

* Corresponding author at: Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Campus Universitário de Santiago, University of Aveiro, 3830-193 Aveiro, Portugal. Tel.: +351 234 370 500.
E-mail addresses: danielpape@ua.pt (D. Pape), lmtj@ua.pt (L.M.T. Jesus), peter.birkholz@tu-dresden.de (P. Birkholz).

http://dx.doi.org/10.1016/j.specom.2015.09.001
0167-6393/

for phonologically voiced English fricatives, but very low for phonologically voiceless fricatives. In contrast to the findings of Stevens et al. (1992), Cho and Giavazzi (2008) observed a linear relation of alveolar fricative voicing in American English listeners, with the result that all listeners showed a linear increase of their voicing responses for increasing vocal fold vibration duration (i.e., for increasing *voicing maintenance*). It has to be noted that when stimuli were presented with only very small portions of fricative voicing (i.e., a few ms) for all listeners a consistent initial bias towards voiced responses was observed, all in comparison to the presentation of fricatives without any voicing. In other words, the linear relationship between stimulus voicing and perception of voicing is not observed for the beginning of the voicing continuum. Rivas (2006) compared Dutch and Italian listeners with respect to their cue-weighting between *frication duration* and *voicing maintenance*. They found that for both languages listeners mainly depended on *voicing maintenance* to discriminate voiced and voiceless alveolar fricatives. Following Rivas (2006), we will use in this paper the term *voicing maintenance* as the percentage of voicing to the length of the entire fricative. This measure is the ratio of the duration of the voiced section to the length of the entire fricative, times 100, expressed in percentage.

Apart from the fricative *voicing maintenance*, both Stevens et al. (1992), Cho and Giavazzi (2008) showed that varying the duration of the fricative noise had a small, but consistent effect on the identification of voicing for American English listeners, with the effect being more pronounced for shorter fricative durations (the shorter the duration, the more the fricative was perceived as voiceless). Denes (1955) showed the same relationship for Dutch intervocalic alveolar fricatives, but Rivas (2006) could only find a very minor effect of fricative noise duration on listeners' identification of voicing for the same place of articulation. With respect to the preceding and following vowel duration, both Denes (1955) and Slis and Cohen (1969) showed that the preceding vowel duration strongly influences the perception of listeners' fricative voicing for Dutch, with longer vowels triggering more voiced fricative responses. Cho and Giavazzi (2008) replicated these results for American English and could confirm the effect of the following vowel duration. In summary, for fricative perception the importance of the three factors (fricative *voicing maintenance*, its *duration* and the preceding *vowel duration*) seems to be comparable to the perception of stops (see, e.g., Pape and Jesus (2014a) for an overview of stop voicing cues in perception). Furthermore, there seems to be both uncertainty and language-dependency with respect to the use of the available acoustic cues for identifying voiced vs. voiceless fricatives (fricative duration, preceding vowel duration or voicing maintenance) and especially how cue-trading between different acoustic cues is mediated by the perceptual system for the language in question.

If the results of Stevens et al. (1992) were valid for all languages, listeners of European Portuguese (EP) would judge the majority of the phonologically voiced fricative productions in the EP database by Pape and Jesus (2015) as voiceless, since most of them are highly devoiced, independent of their place of articulation and vowel context. In other words, if the presence of voicing during the fricative portion of the stimulus was the main factor for voiced/voiceless distinction, then most of EP voiced items would be judged voiceless. Thus, apparently additional cues take over (separately or combined) to result in a robust perceptual voicing distinction, and are weighted against the maintenance of fricative voicing. This discrepancy between the phonetic realisations of the EP production data and the expected perceptual responses based on Stevens et al. (1992) and other studies lead directly to our research questions, i.e., whether (1) perceptual ratings in fact mirror the very low voicing maintenance of EP phonologically voiced fricatives in speech production (Pape and Jesus, 2014b, 2015; Jesus and Shadle, 2002; Jesus and Shadle, 2003b; Pinho et al., 2012) and (2) cue-weighting between acoustic cues like preceding *vowel duration*, *fricative duration* and fricative *voicing maintenance* would occur for EP fricative perception. The second research question is particularly difficult to answer due to findings that cue-weighting is listener-dependent, i.e., some listeners give more weight to certain acoustic cues than others. Furthermore, redundant cues often enter into trading relationships; thus, increasing the amount of certain cues while at the same time decreasing the amount of other cues may generate the same perceptual outcome.

In the present study, we performed perception experiments to assess these two questions. In general, methods to define and build a multidimensional stimuli space for perceptual experiments should be able to generate realistic articulatory targets, while simultaneously allowing independent parametric control of perceptual parameters, such as fricative duration, vowel duration and transition duration. The most realistic perceptual stimuli would consist of naturally recorded speech, but in this condition there is no control of the presence and strength of possible perceptual cues, thus introducing an unwanted perceptual bias to the experiment. On the other hand, articulatory synthesis can generate realistic articulatory targets and the independent parametric control of all relevant parameters, but suffered until recently from very poor noise generation capabilities. For our study, we overcame this drawback by using an articulatory speech synthesiser with a sophisticated noise source model for stimuli generation, where all relevant parameters of the stimuli could be controlled at a physiological level (Birkholz, 2013, 2014; Birkholz et al., 2011; Birkholz and Jackèl, 2004).

Furthermore, using an articulatory synthesiser allowed us to examine the perceptual relevance of two additional differences in the production of voiced vs. voiceless fricatives: Studies using EMMA and MRI data have shown that in sustained productions there are no significant articulatory differences with respect to the place of the main articulator between voiced and voiceless fricatives in