Constructing Physically Realistic VCV Stimuli for the Perception of Stop Voicing in European Portuguese

Daniel Pape[1], Luis M.T. Jesus[1,2], and Pascal Perrier[3]

[1] Institute of Electronics and Telematics Engineering of Aveiro (IEETA), University of Aveiro, 3810-193 Aveiro, Portugal
[2] School of Health Sciences (ESSUA), University of Aveiro, 3810-193 Aveiro, Portugal
[3] DPC/Gipsa-Lab, UMR CNRS 5216, Grenoble INP, Grenoble, France
{danielpape,lmtj}@ua.pt,
{Pascal.Perrier}@gipsa-lab.grenoble-inp.fr

Abstract. In this book chapter we present the generation of physically realistic stimuli with a biomechanical speech production model, with the aim to produce perceptually appropriate VCV sets for the European Portuguese (EP) voicing distinction. The duration measures necessary for the biomechanical model were extracted from an extensive EP speech production database, recorded for this aim. The same database was used to generate realistic voicing extinction contours for the perceptual continuum. To assess the realistic accuracy of the biomechanically generated stimuli, we compared the biomechanical stimuli set to linear interpolation between articulatory targets, traditionally used for speech synthesis.

Keywords: biomechanical modelling, perceptual cues, cue weighting, European Portuguese, voicing perception.

1 Introduction

The work outlined in this book chapter consists of perceptual stimuli modelling as part of a research project on the importance of *voicing maintenance* in both speech production and perception in European Portuguese (EP) compared to other languages. For velar stop perception, we used and compared extracted voicing patterns and durational values from real speech productions (see Pape & Jesus 2011) in a matched cross-linguistic speech perception study, with the aim to examine the actual use and interaction (cue weighting) of the perceptual cues vowel duration, consonant duration and voicing maintenance. The speech material generated for the perceptual experiments consisted of biomechanically modelled stimuli acoustically synthesized with a three mass vocal fold model. The biomechanical modelling has the main advantage that all obtained tongue movements, trajectories and phoneme targets are comparable to natural speech, but with the additional possibility to manipulate all important temporal and glottal source parameters while maintaining articulatory realism. In sum, the use of biomechanical modelling is the best compromise to guarantee highly realistic perceptual stimuli, and to independently control parameters such as duration, transition and targets.

H. Caseli et al. (Eds.): PROPOR 2012, LNAI 7243, pp. 338–349, 2012.

1.1 Perceptual Cues for Stop Voicing

For speech production, phonological voicing distinction is defined as the presence or absence of vocal fold vibrations during consonant production (Jakobson et al. 1952). For speech perception, a stop voicing distinction is mainly based on Voice Onset Time (VOT) (Lisker & Abramson 1964, 1967). Cross-linguistic differences in voicing perception are captured by changes in the VOT boundaries, i.e., by the location of the identification boundaries between voicing categories on a VOT continuum (Hoonhorst et al. 2009). In languages with three voicing categories (voiced, voiceless and voiceless aspirated) the mean VOT boundaries are located around ±30ms. Two-category languages differ in the nature of their voicing categories. In languages with a voiceless aspirated contrast the boundary is at +30ms (Lisker & Abramson 1970), whereas for languages with voiced/voiceless contrast without aspiration the boundary is 0ms (for Spanish: Williams 1977; for French: Serniclaes 1987). Infants below six months of age raised in an English environment are sensitive to both VOT boundaries (±30ms, 0ms), although only the positive VOT boundary is phonological in English (Aslin et al. 1981) or the 0ms VOT in the other language (French: Hoonhorst et al. 2009; Spanish: Lasky et al. 1975).

VOT is one of the most dominant cues for characterising stops in a number of languages, but a number of additional perceptual cues are found to influence the perception of voicing: consonant and adjacent vowel duration (Luce & Charles-Luce, 1985; Jessen 1998; Cuartero 2002; Viana 1984), and loudness (Repp 1979), among others. These other cues distinguish voicing, in combination with VOT but also without VOT, i.e., when VOT is ambiguous or missing. Further, the literature shows (Morrison 2005; Escudero et al. 2009) that 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. Taking into account the variety of perceptual cues for stop voicing, the question arises how different languages weight the available cue to achieve robust perception. However, few studies (Francis et al. 2000) attempted to study the simultaneous variation and cue weighting for stop voicing distinction, and (to our knowledge) no studies examined this cue weighting framework for stop voicing in a cross-linguistic context. The last point is of utmost importance when one takes into account speech production differences, for example, in voicing maintenance between different languages (see, e.g., Solé 2011 for Spanish vs. English, and Pape et al. (submitted) for EP vs. German and Italian). Given these cross-linguistic differences, the question arises whether these are also reflected in the perception, and if so, how these differences influence the cue weighting of the available cues.

1.2 Modelling Physically Realistic Perceptual Stimuli

One important point to take into account when designing a valid multidimensional cue weighting experiment are the transitions between the phoneme targets: For example, velar stops show a strong articulatory forward loop (see, e.g., Mooshammer et al. 1995), additionally the shape of the loops differs for voiceless and voiced consonantal targets (Brunner et al. 2011). Assuming that these loops could play a perceptual role, in the design of multidimensional perceptual stimuli one preferably