Does physical realism of articulatory modelin improve the perception of synthetic speech?

Daniel Pape, Pascal Perrier, Susanne Fuchs, Sonia Kandel

Does physical realism of articulatory modeling improve the perception of synthetic speech?

Daniel Pape[1,3], Pascal Perrier[1], Susanne Fuchs[2], Sonia Kandel[4]
[1]GIPSA-lab, CNRS UMR 5216, Grenoble University, France
[2]Phonetik/ZAS, Berlin, Germany
[3] IEETA, Aveiro University, Portugal
[4]LPNC, CNRS, Grenoble University & Institut Universitaire de France, France

1. INTRODUCTION

The role of the precise time variation of formants (henceforth formant trajectories) in the perception of speech has been at the center of many recurrent debates, essentially along two opposing theories. For the first theory, formant trajectories do not contain any relevant phonetic information by themselves. In the best case, when properties other than the starting and ending points matter [1], the trajectories would only be the vector of information related to the intended target [2]. For the second theory, the temporal characteristics of the transition would be, independently of any target, the relevant perceptual information. Thus, formant trajectories in all their details would be the object of perception ([3]). Evidence supporting this theory has been recently provided by the perturbation study carried out by Cai et al [4]. In presence of a real-time perturbation of their perceived formant trajectory, Chinese speakers of Mandarin producing the triphthong /iau/ tended to react by modifying their articulation in order to generate the usual trajectory.

A possible way to understand the different findings made in the context of both main theoretical streams is inspired by the work of Viviani and colleagues (i.e. [5]) on visual perception and the identification of hand gestures. These authors have found evidence that this perceptual task could be strongly determined by the knowledge that human beings have of the physical characteristics of their own movements. In this context, it can be assumed that formant transitions that are not compatible with the physical mechanisms underlying speech production should be perceived as being incorrect. This could explain compensation strategies such as those observed by Cai et al. In normal speech, the physical mechanisms of speech production could determine the patterns of the spectro-temporal variation between discrete targets, and these patterns would in turn become "the" physical objects associated with the perception of sound sequences.

This article presents the first step in a process of evaluating the potential impact of the physical properties of the articulators on the perception of speech. Perceptual tests of synthetic stimuli generated with different models incorporating various degrees of physical complexity of these articulators have been run and analyzed.

2. SUBJECTS AND METHOD

Subjects

Twenty-three subjects (16 men, 7 women, aged between 25 and 50 years) participated in the experiment. None of them was aware of the methods used to generate synthetic speech. All are native speakers of French. None of them has reported any speech or hearing problem.

Method

Vowel1-Vowel2-Vowel1 (V1V2V1) and Vowel1-/g/-Vowel1 (V1CV1) acoustic stimuli were synthesized from sagittal vocal tract shapes, via the generation of an area function and the use of a Kelly-Lochbaum acoustic model ([6]). In all cases, the fundamental frequency was kept at 110Hz. The three different models used to generate the stimuli differed in how the transitions between target vocal tract shapes are generated. In the first class of stimuli (Mod1) transitions were obtained with a two-dimensional biomechanical model of the vocal tract [7] controlled on a target basis. Motor commands are specified for each elementary sound and movements are generated by a time shift of the target commands at a constant rate [7]. For Mod2 and Mod3 stimuli, the target vocal tract shapes are the ones actually reached in Mod1 stimuli for V1, V2 and C. The timing was the same for all classes of stimuli. For Mod2 and Mod3 the transitions between the target shapes were computed along straight paths. These two models differ in the time course of the displacements along these path. In model 2 (Mod2), the displacement has a constant speed. In model 3 (Mod3),

speed is an arc of sinusoid in line of the kinematic properties of an undamped second order system. In sum, Mod1 stimuli correspond to the most realistic physical model, Mod2 stimuli are less realistic and Mod3 are physically the least realistic ones. For another perception test, degraded stimuli V1V2V1 were also generated by replacing the central part of V2 by a silence with the same duration according to the paradigm of the silent centers [3].

All perception tests were conducted in an anechoic chamber at GIPSA-lab. The subjects were first asked to assess the silent center stimuli. The task of the listeners was to identify V2. A choice was given between 5 possible vowel answers. Then subjects were asked to assess the naturalness of the non-degraded stimuli (V1V2V1 first and V1CV1 afterwards). All experiments were forced choice. The reaction time (RT henceforth) was measured.

The purpose of the statistical analysis is to see if, for all subjects taken together, there are differences between the Mod1, Mod2 and Mod3 stimuli. We used the generalized linear model with mixed effects. The class of stimuli (Mod1, Mod2, Mod3) was chosen as the fixed factor.

3. RESULTS

These tests assessing the potential link between the degree of realism of our models and the perceived naturalness of the synthesis did not show any significant results. As concerns the silent centers stimuli, with respect to the percentage of correct responses given for all the subjects, no difference between the models can be shown. However, various studies have shown that the listeners' Reaction Time should be considered when analyzing these data. Indeed, long Reaction Times may be related to the involvement of high-level cognitive processing, which are beyond the scope of our study, namely the auditory perception itself. Hence, in a second analysis, only the responses (all subjects taken together) corresponding to short Reaction Times in the [1000ms 2000ms] interval were taken into account. Under this condition, the identification of the target vowel V2 in the silent center experiment was significantly better for the Mod1 stimuli than for the Mod2 stimuli. A trend was also observed suggesting a better identification for Mod3 stimuli as compared to Mod2 stimuli. The differences between Mod1 and Mod3 are not significant.

4. CONCLUSIONS

As concerns the perceptual rating of the naturalness of the three classes of stimuli, our study did not show any influence of the degree of realism of the underlying articulatory models. This could be due to the fact, that even in the best cases the acoustic synthesis is still far from being equivalent to natural speech (among other because of the constant F0). The silent centers experiments have shown that the stimuli generated with a pure kinematic model, integrating no biomechanical and no dynamic influences allowed a significantly worse identification of the missing central vowel V2. It can be interpreted as evidence for the fact that information related to the dynamics of the vocal tract articulators is used by listeners to recover the intended but not uttered vowel target.

[1] Sussman, H.M., Fruchter, D., Hilbert, J. & Sirosch, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *BBS, 21***,** 241–299

[2] Lindblom, B. & Studdert-Kennedy M. (1967) On the role of formant transitions in vowel recognition. *JASA, 42,* 830-843.

[3]Strange, W., Jenkins, J.J. & Johnson, T.L. (1983). Dynamic specification of coarticulated wovels, *JASA, 74(3),* 695-705.

[4] Cai, S., Boucek, M., Ghosh, S.S., Guenther, F. H. & Perkell, J.S. (2008). A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iau/. *Proc. of ISSP-2008.*, (pp. 65-68), Strasbourg, France.

[5] Viviani, P., & Stucchi, N. (1992). Biological movements look uniform: evidence of motor-perceptual interactions *J Exp Psychol Hum Percept Perform*, 18 (3), 603-623.

[6] Story, B.H. (2005). A parametric model of the vocal tract area function for vowel and consonant simulation, *JASA, 117(5),* 3231-3254.

[7] Perrier, P., Payan, Y., Zandipour, M. & Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: a modeling study. *JASA, 114 (3),* 1582-1599.