DYNAMIC VERSUS STATIC SPECIFICATION FOR THE PERCEPTUAL IDENTITY OF A COARTICULATED VOWEL

Michel Pitermann

Laboratoire Parole et Langage, ESA 6057 CNRS
Université de Provence, 29 Ave. Robert Schuman
13621 Aix-en-Provence Cedex, France
mpiter@lpl.univ-aix.fr

ABSTRACT

This paper presents a perceptual experiment on stimuli synthesized by means of a vocal tract area function model. The purpose was to compare the contribution of dynamic against static information to the identity of a coarticulated vowel. Three sources of information were perceptually analyzed: (i) the vowel nucleus; (ii) the acoustical contrast between the vowel nucleus and the "stationary" parts of its immediate context; (iii) and the transitions linking the stable parts of the speech signal. The results show that the vocoïds were better identified by dynamic information. This backs up the perceptual overshoot model proposed in Lindblom and Studdert-Kennedy (1967). However, this conclusion must be confirmed by further experiments.

1. INTRODUCTION

The nature of information perceptually processed to identify a vowel in continuous speech is still debated at length (see Nearey (1989) and van Son (1993) for two discussions). Three kinds of information are often analyzed: (i) vowel intrinsic information contained in its nucleus (Assmann et al. 1982); (ii) the acoustical contrast between the vowel nucleus and other stationary parts of the speech signal (Ainsworth 1975); (iii) and dynamic information contained in the transitions linking these stable parts (Strange et al. 1983; Strange 1989). Although these three kinds of information seem to contribute to the perceptual identity of a vowel, it is not clear which one is predominant in continuous speech (Nearey 1989).

Some perceptual experiments on speech signals synthesized by means of a vocal tract model were presented in an ASA meeting (Carré et al. 1994). One of them consisted in comparing the vowel identification scores measured for sustained isolated vocoïds and for stimuli containing only intervocalic transitions. The synthetic vocoïds containing dynamic information were better categorized. Hence, the authors' conclusion was that the subjects had used articulatory gesture information encoded in the transitions. However, some evidence that a static context can modify the perceptual categorization of a sustained vocoïd had been shown (Nearey 1989). As a consequence, the vowel context might suffice to explain the higher identification scores measured for the vocoïds containing transitions in the Carré et al. (1994) investigation. Therefore, we extended their experiment to analyze this hypothesis.

2. METHOD

The method consisted in synthesizing vocoïds in a vowel continuum by means of a vocal tract area function model. The stimuli contained either static or dynamic information. Nine naive subjects were asked to categorize them in a perceptual assignment. The position in the formant space of the perceptual boundary separating two vowel categories was analyzed as a function of the type of stimuli synthesized. It indicates which kind of information was predominant for the perceptual identity of the synthetic coarticulated vocoïds.

The vocal tract area function model was a six-tube Kelly-Lochbaum model (Schoentgen and Ciocea 1995b). Five section areas out of six and the total length were the model parameters.

Time series of the model parameters were estimated from a natural speech signal [iai] by means of an acoustic-to-geometric inversion (Schoentgen and Ciocea 1995a). Three types of stimuli were built from these time series: (i) isolated sustained vocoïds ([V] signals); (ii) the same sustained vocoïds presented between two sustained [i] ([i#V#i] signals, where # stands for a 20-ms silence); (iii) and vocoïds containing only [ia] and [ai] transition segments ([iVi] signals). The details are described in Figure 1.

The complete [ia] transition contained 12 samples of the model parameters. Therefore, three sets of 12 stimuli were built. Each of them was replicated six times in order to create a 72-item list of isolated sustained vocoïds [V] and a 144-item list of mixed [i#V#i] and [iVi] stimuli.

To synthesize an acoustic signal from the time series of the six model parameters, the corresponding values of the first three formants were calculated for

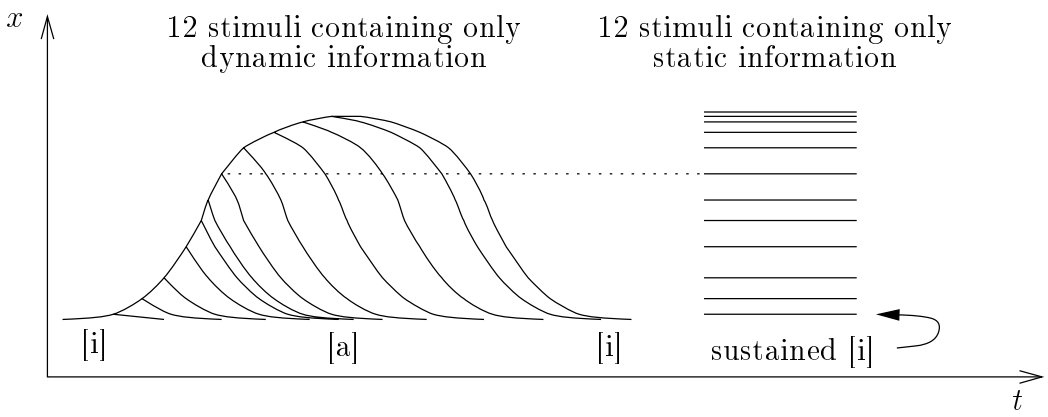

Figure 1: Vocoïd construction. The graphic shows a fictitious feature taking a high or low value respectively for a [a] or [i] production. To synthesize a sustained vocoïd, the values of the model parameters estimated at a time coordinate were duplicated for a 300-ms period (stimuli shown on the right-hand side of the graphic). The sustained [i] used in the [i#V#i] stimuli was synthesized in the same way, i.e. copying the first vector of the model parameters estimated in the [ia] transition. To build a symmetric [iVi] signal containing only transitions, a mirror image of a [ia] transition segment was appended to the ordinary one (stimuli shown on the left-hand side of the graphic). The dotted horizontal line shows a pair of corresponding [iVi] and [V] stimuli.

each time coordinate (Schoentgen and Ciocea 1995b). The values of the fourth and fifth formant frequencies were kept constant respectively at 3250 and 3700 Hz. Then, the acoustic signal was synthesized on the version 3.04 implementation of the Klatt synthesizer with a sampling rate of 16 kHz (Klatt 1980; Klatt and Klatt 1990). To avoid onset and offset noise, the signal intensity was multiplied by a 12.5-ms linear slope at the beginning and end of each stimulus.

Nine French speaking naive subjects, aged between 22 and 30, listened to the two stimulus lists. None reported suffering from impaired hearing. The synthetic signals were shuffled for each participant in order to avoid biased identifications due to the stimulus order of presentation. Five listeners started with the [V] list, the other four with the second list. Two successive items were separated by a 3-second silence. Before listening to his list, each subject could hear 20 items to adjust the signal intensity to a comfortable level and to become familiar with the assignment.

The task was to write down each vowel heard for the [V] list or each vowel identified between the two [i] for the other list. When the sound was not deemed a vowel, the listener could leave a blank or choose the closest vowel. No information was given about the vowels synthesized or phonetic transcription to use, therefore it was a test with open responses.

The signals were played back on BEYER Beyerdynamic dt325 headphones connected to a Sparc 20 SUN workstation containing a 16-bit linear digital to analog converter.

3. RESULTS

The subjects identified the vocoïds of the [a]–[i] continuum as [a], [ə], [ɛ], [e] or [i]. Some stimuli were labeled [ə] because the [ia] transition passed through the point $(500, 1650, 2470)$ Hz in the (F_1, F_2, F_3) space. However, some participants reported that the [ə] quality was particularly poor.

Figure 2 shows the [a] identification score as a function of the first formant. The first formant value of an [iVi] stimulus was defined as the highest value used in the synthesis, i.e. the value used to synthesize the corresponding sustained vocoïd (see Figure 1).

The results were homogeneous with respect to the nine subjects. The only exception was the set of the three sustained isolated vocoïds [V] corresponding to the three highest values of the first formant (the three right '$*$' signs of Figure 2). For these stimuli, the listeners were divided into two groups: five participants systematically classified the three vocoïds as [ə], the other four often labeled them [a] producing approximately the same identification scores for the [V] and [i#V#i] signals.

Figure 3 shows in the (F_1, F_2) space the natural [iai] speech signal used to build all the synthetic stimuli.

4. DISCUSSION

The results presented in Figure 2 show that the isolated sustained vocoïds [V] were poorly categorized as [a]. Hence, the nucleus of the coarticulated [a] did not correspond to a good sustained [a]. The values of its first three formants were 630, 1500 and 2400 Hz.

Figure 2 shows also that the static [i#_#i] context modified the perception of the synthetic sustained vocoïds. The same conclusion was found in Nearey (1989) for [i#ɒ#V] stimuli, when two different pairs of [i,ɒ] were used as two similar but different contexts.

The best identification scores were obtained here for the [iVi] stimuli. Six scores out of 12 were higher

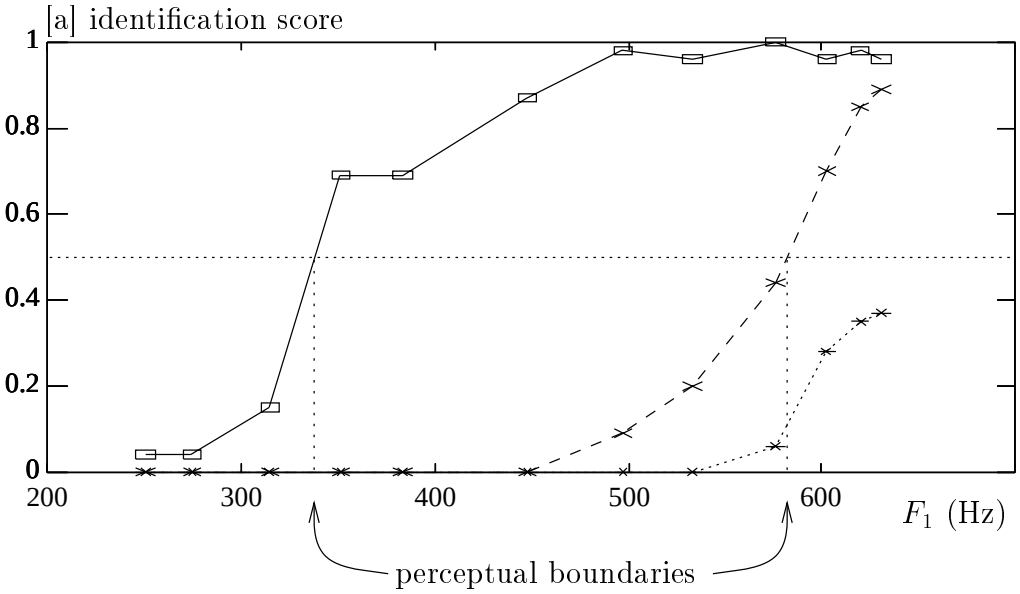

Figure 2: [a] identification score as a function of the first formant. The '□', '×' and '∗' signs represent the results respectively for the [iVi], [i#V#i] and [V] stimuli. Each perceptual boundary between the [a] and [ə] categories was estimated by means of a linear interpolation between the two scores bracketing the 0.5 value as shown by the dotted horizontal and vertical lines.

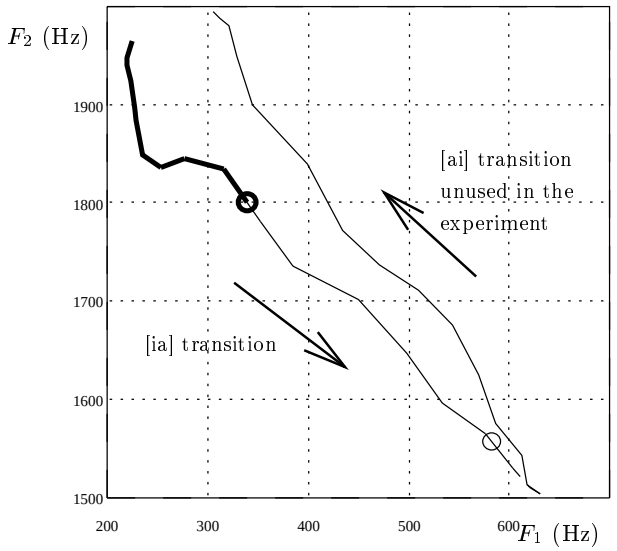

Figure 3: Representation in the (F_1, F_2) space of the original [iai] speech signal used to build the synthetic ones. The thick and thin ellipses show the perceptual boundaries between the [a] and [ə] categories respectively for the [iVi] and [i#V#i] stimuli. The thick line is the formant excursion of the synthetic [iVi] stimulus needed to reach the 0.5 identification score.

than 0.95. When the values of the first three formants were 450, 1700 and 2500 Hz, 87 % of the [iVi] signals were still categorized as [a]. In contrast, no corresponding sustained vocoïd was labeled [a]. The smallness of the formant excursion needed to perceive a [a] in a [iVi] stimulus can be seen on Figure 3.

The mean difference between the identification scores measured for two kinds of stimuli was estimated by:

$$\sqrt{1/12 \sum_{i=1}^{12} (y_i - x_i)^2}, \qquad (1)$$

where y_i stands for the identification scores measured for the first type of stimuli and x_i for the other ones. This estimate led to 0.27 for the difference between the scores measured for the [i#V#i] and [V] stimuli and to 0.54 for the [i#V#i] and [iVi] signals. It shows that the vocoïds in the middle of the [i#V#i] stimuli were perceptually closer to the [V] than [iVi] signals. Hence, the vocoïd context can explain only a small part of the difference between the scores measured for the [iVi] and [V] signals. Therefore, *the perceptual identity of the synthetic coarticulated vocoïds depended mainly on the dynamic of the model parameters.*

In the (F_1, F_2) space, the perceptual boundary between the [a] and [ə] categories was estimated at (340,1800) Hz for the [iVi] stimuli and (580,1560) Hz for the [i#V#i] (their positions are shown on Figure 3). Therefore, the boundary shifts from the [iVi] to [i#V#i] stimuli were +240 Hz for F_1 and -240 Hz for F_2. It amounts to 340 Hz in the (F_1, F_2) space. A boundary shift can be interpreted as a perceptual

overshoot, i.e. a perceptual extrapolation of a formant transition beyond the value actually reached (Lindblom and Studdert-Kennedy 1967). Therefore, *our results bring support to the perceptual overshoot model.*

In van Son (1993), the author made the hypothesis that a perceptual overshoot occurs when the vowel context has been identified. With an unidentified context, the listener would categorize a vowel token by means of a weighted formant average. Our results are compatible with this hypothesis because a perceptual overshoot was observed for the [iVi] stimuli and the subjects were told what the vowel context was, so it was identified. However, our experiment cannot really back or contradict van Son's hypothesis because the vowel context of the [iVi] stimuli was always identified. To study the hypothesis, we should add vocoïds with the same transitions but an unidentifiable context. The comparison between the perceptual boundaries between the vowel categories estimated for these signals, the [iVi], [i#V#i] and [V] stimuli would indicate if a perceptual overshoot, or a weighted formant average or a third possibility would have occured.

We have argued so far that our synthetic vocoïds were better identified by dynamic than static information. These results also bear out the perceptual overshoot hypothesis. However, they are the fruit of one experiment involving a single vowel continuum with one vocalic context. As a consequence, the conclusion must not be generalized.

We have not specified yet the nature of dynamic information perceptually involved to categorize the synthetic vocoïds. This experiment was not designed to answer that question. As previously stated, the time series of the model parameters were estimated by means of an acoustic-to-geometric inversion applied to a natural speech signal. Thereby, no assumption was made about the dynamic control of the model parameters. To study that issue, we should define and perceptually compare several kinds of dynamic control.

5. CONCLUSION

The contributions of dynamic and static information to the perceptual identity of a coarticulated vowel were compared in an experiment carried out with stimuli synthesized by means of a vocal tract model. The results show that the perceptual identity of the synthetic coarticulated vocoïds depended mainly on the dynamic of the model parameters, but the nature of the dynamic involved is still unknown. These results bear out the perceptual overshoot hypothesis. However, owing to the smallness of the corpus, the conclusion must not be generalized.

ACKNOWLEDGMENTS

This research was carried out in the Institute of Modern Languages and Phonetic in Bruxelles. Thanks are due to Sorin Ciocea for carrying out the acoustic-to-geometric inversion used in this work.

REFERENCES

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant and M. Tatham (Eds.), *Auditory Analysis and Perception of Speech*, pp. 103–113. Academic, London.

Assmann, P., T. Nearey, and J. Hogan (1982). Vowel identification: Orthographic, perceptual and acoustic aspects. *The Journal of the Acoustical Society of America 71*, 975–989.

Carré, R., S. Chennoukh, B. Lindblom, and P. Divenyi (1994). On the perceptual caracteristics of "speech gestures". *The Journal of the Acoustical Society of America 96*, pp. 3326 (Abstract). ASA meeting held in Austin.

Klatt, D. and L. Klatt (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America 87*(2), 820–857.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America 67*(3), 971–995.

Lindblom, B. E. F. and M. Studdert-Kennedy (1967). On the role of formant transitions on vowel recognition. *The Journal of the Acoustical Society of America 42*(4), 830–843.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America 85*(5), 2088–2113.

Schoentgen, J. and S. Ciocea (1995a). Direct calculation of the vocal tract area function from measured formant frequencies. In *Eurospeech'95 Proceedings*, Volume 1, Madrid, Spain, pp. 745–748. European Speech Communication Association.

Schoentgen, J. and S. Ciocea (1995b). Kinematic acoustic-to-geometric mapping. In K. E. . P. Branderud (Ed.), *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Volume 2, Stockholm, Sweden, pp. 194–197.

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *The Journal of the Acoustical Society of America 85*(5), 2135–2153.

Strange, W., J. J. Jenkins, and T. L. Johnson (1983). Dynamic specification of coarticulated vowels. *The Journal of the Acoustical Society of America 74*(3), 695–705.

van Son, R. J. (1993). Vowel perception: a closer look at the literature. In *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, Volume 17, pp. 33–64.