SPEECH DYNAMICS: EPISTEMOLOGICAL ASPECTS

René Carré, François Pellegrino*, Pierre Divenyi***

*Laboratoire Dynamique du Language (DDL), CNRS-Université de Lyon 2
**VA Medical Center and EBIRE, Martinez, California
recarre@wanadoo.fr; francois.pellegrino@univ-lyon2.fr; pdivenyi@ebire.org

ABSTRACT

Speech is generally looked upon as a succession of events in the time domain and analyzed frame by frame, while ignoring the fact that speech is dynamic. In the present paper, evidence in support of the dynamic nature of speech and dynamic invariance, as well as their consequences on speech research, are discussed.

Keywords: speech production, speech dynamics, speech kinematics.

1. INTRODUCTION

The study of speech communication is rooted in a long-lasting paradox: Alphabetic writing systems have been developed because speech can be perceived as a temporal succession of a relatively small number of elementary sounds, called phonemes. We termed phonemes as abstract and described without taking into account their temporal aspects because spectrographic analysis reveals that it is impossible to segment the speech signal into discrete phonemes. Characteristics of phonemes also display considerable variability across speakers and phonemic environment (co-articulation [1,2] and vocalic reduction [3], etc.). Moreover, numerous studies show that speech production is syllabic (i.e., is a succession of syllables obtained by co-production of consonant and vowel [4,2]) and ill-described in static terms. There are no invariant acoustic measurements that could really be used to characterize phonemes: even standard formant targets used for vowel description cannot be regarded as invariant since they exhibit significant variations across speakers – male, female, child subjects – and languages. Plosive consonants are described by their loci [1] or their locus equation [5], resulting only in relational invariance. Burst characteristics are also used to describe consonants [6]. It is clear that these descriptions are almost exclusively static and spectral, leaving out the temporal dimension for the characterization of phonemes.

However, vowel perception can be definitely improved by taking temporal information into account [7,8]. "Silent center" experiments have clearly shown that perception of vowels [9] as well as consonantal places of articulation [10] use information from adjacent transitions. It was also observed that taking into consideration two formant values on either side of a vocalic target improves vowel recognition [8]. Also, Furui [11] demonstrated the importance of the dynamic cepstral coefficients in speech recognition.

In spite of these results, speech analysis and speech theories continue to regard the speech signal as a succession of time-windowed frames (e.g., [12]), although it is clear that such a static approach fails to properly acknowledge the fundamentally dynamic aspect of formant evolution and the intrinsic temporal characteristics of speech sounds [13]. In short, the fact that most speech theories can still be qualified as static [14], makes it imperative to stress the necessity of a dynamic alternative.

Our objective here is to emphasize the intrinsically dynamic (taken here in its narrow sense of kinematics) nature of speech by examining voiced formant frequencies and show the epistemological consequences that follow.

2. TOWARDS A DYNAMIC APPROACH

If speech were nothing but a succession of events that occur at specific points of time – such as spectral extrema or onsets/offsets of transitions – then, in order to accurately characterize the events, precise measurements must be accessible at these specific instants. However, it is common knowledge that it is difficult to precisely determine the temporal position of the locus and, consequently, also that of the corresponding formants. It is similarly difficult to measure the formant frequencies for high-fundamental-

frequency voices (female and children's voices). Such measurements become outright impossible in a noisy or reverberating environment.

Furthermore, the phenomenon of vowel reduction assumes that, due to the high degree of inertia of the articulators, the articulatory mechanism is unable to reach the intended targets. If reaching these targets were an absolute requirement, as many static theories imply, it would mean that the speech production mechanism is ill-adapted to perform an everyday task. Since this conclusion is not acceptable, we may assume instead that the task does *not* consist of reaching static targets. Several studies that support the dynamic nature of speech appear to answer this question affirmatively.

2.1. Functional arguments

Let us suppose that speech production is the result of an emergent evolutionary process. If this were the case, then reaching a given static target (that will be only known when the evolution is complete) could not be the goal during the process itself. Rather, one possible goal may be the increase of the acoustic space by means of an increased displacement of the articulators. Positive feedback is likely to reinforce these displacements, considered as movements *from* an origin *along* specific directions, consequently without any need for static targets.

Moreover, the speech communication system is dynamically optimal [15] if:

- a minimum deformation of the tube leads to a maximum of acoustic variation (more or less equivalent to a minimum effort criterion);
- the sequence of coding gestures is not represented in terms of a succession of static parameters but in terms of variations (as in delta modulation), thus minimizing the amount of information to transfer;
- the amount of information transfer is increased by parallel transmission of different coding gestures.

2.2. Syllabic co-production

Studies demonstrating that production is syllabic [4] effectively lead to production dynamics and gestural approach which, incidentally, was at least suggested by articulatory phonology [16]. At a specific point in time, abstract commands corresponding to a syllabic CV will evoke, at the peripheral level, co-production of consonant gestures superimposed on vocalic gestures [2].

Obviously, gestures, just like any vector, can be characterized by their starting points and targets: in this case, they are obtained by interpolating between the two endpoints. Oddly, the outcome of this is scarcely more than an extension of the static approach. But, a gesture can also be characterized by its starting point and direction (in the articulatory or acoustic spaces) and by its velocity (of deformation in the articulatory space or displacement in the acoustic space), giving rise to a true dynamic approach where targets are not defined in the acoustic or articulatory space but in their derivatives with respect to time.

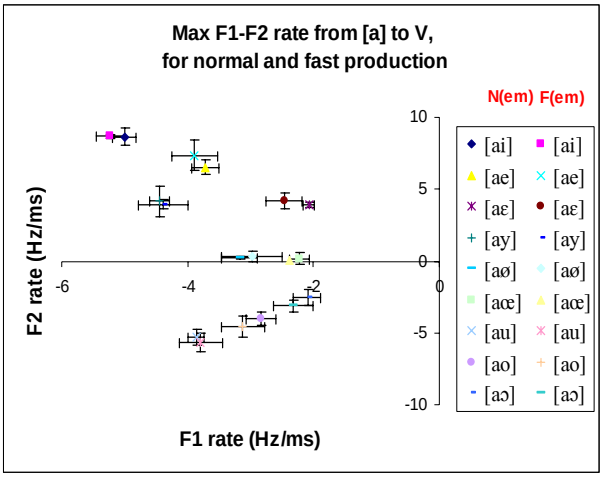
3. DYNAMIC INVARIANCE

If speech representation is characterized by dynamic parameters then it can be assumed that at the production level these parameters must be invariant to a certain degree and that the same parameters are also used at the perception level. This section exposes results from two preliminary experiments assessing this invariance.

3.1. Formant transition direction and rate in the F1-rate/F2-rate plane

Figure 1 shows the formant transition rates (means and standard deviations for five [aV] productions by one speaker) in the F1 rate/F2 rate space, at normal and fast speech rates [17]. V is one of the French oral vowels. The rates retained are the maximum rates of the transitions.

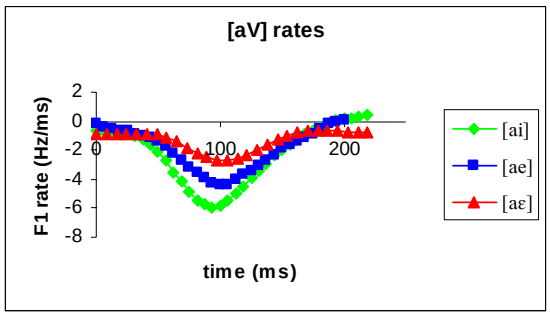
Figure 1: Vowel transition maximum rates (and standard deviation) in the F1 rate/F2 rate space of the transition [aV] for normal (N) and fast speech rates (F) (5 productions of speaker em).



It can be observed that the standard deviation for F1 is larger than for F2. A possible explanation is that the formant measurement procedure gives an absolute error which is relatively more important at low frequencies.

The rates depend on the position of the starting point (vowel [a]) in the speaker's F1-F2 plane. We do not observe large differences between normal and fast production and the overlap between vowels mainly affects trajectories from [a] to back vowels. Moreover, vowels can be described *dynamically* starting at the very beginning of the transition. Figure 2 shows three transition rates (for [ai], [aɛ], [ae]) synchronized at the beginning. The three vowels can be distinguished based on the maximum rates corresponding more or less to the middle of the transition (max. rate of [ai] > max. rate of [ae] > max. rate of [aɛ]) but also all along the transition — thanks to the different slopes pertaining to each vowel. This dynamic approach can explain results by Chistovich [18] on listeners shown to perceive the syllable before they hear it completely and those by Strange on the silent center. Our results suppose that the transition durations are more or less constant for all the [aV], which is the case here (see Fig. 2). Though this observation is consistent with other reports [19,20] and [21] for normal and fast production, it remains to be confirmed with data from more speakers [22,23].

Figure 2: F1 rates in the time domain for [ai], [ae], [aɛ] for speaker (em) at normal rate.



3.2. Dynamic perceptual invariance

In order to assess the dynamic invariance hypothesis, perceptual experiments have been performed with $\mathrm{V}_1\mathrm{V}_2\mathrm{V}_1$ transitions, where V_1 and V_2 are synthesized with formants values outside the traditional F1/F2 vowel triangle [17]. Results show that such transitions can be categorized as vocalic trajectories according to the direction and rate of the transition. For example, a $\mathrm{V}_1\mathrm{V}_2\mathrm{V}_1$ trajectory more or less parallel and equal in size to [iui] of the vowel triangle is perceived as /iui/ though the perceived /u/ is acoustically placed at the [a] of the vowel triangle. A $\mathrm{V}_1\mathrm{V}_2\mathrm{V}_1$ trajectory more or less parallel and equal in size to [aua] is perceived as /aua/ though the perceived /u/ is acoustically also placed at the [a] of the vowel triangle and a shorter in size trajectory is perceived as /aoa/ though the perceived /o/ is placed at the [a] of the vowel triangle. These results, incompatible with a static target approach, support a dynamic approach which takes for granted that humans are able to cope with these derivative – or velocity – parameters. In fact, consistent with our perceptual results, existence of velocity (and acceleration) detectors in the auditory system has been demonstrated in psychoacoustic studies [24,25].

4. DISCUSSION

In the static vowel-target approach, identification occurs only after the transition is completed, thus requiring a backward-integrative procedure to happen. In contrast, in the dynamic approach, identification may occur already at the beginning of, and all along the transition, using knowledge of the point of departure and trajectory direction in the acoustic space. This approach can consequently be termed forward-derivative. This point of departure can be acoustically known but to explain the results of the perception experiment described above, it seems that the point of departure must be phonologically identified. Further studies must be undertaken on the characteristics of this departure point (Absolute characteristics? Relative? Phonetic? Phonologic? Both?).

Further studies are also necessary to assess whether dynamic parameters display a stronger invariance across male, female and child speech than the well-known variability of static vowel targets. The issue of normalization, which operates in the frequency domain in a static approach and in the time domain in the dynamic also approach needs further experiments to decide.

Also, standard techniques used to detect formant frequencies are notoriously ill-suited for the analysis of speech with high fundamental frequencies or in low SNR, and they are not well-adapted for measuring spectral variations. The dynamic approach will thus be compelled to reconsider analysis techniques and to also include

information on the phase of the speech signal in addition to its amplitude, in a way consistent with contemporary auditory theory. Phase variation could become a tool for measuring the rate of the transitions. See also for example [26] where an auditory model capable of detecting spectral transition without formant tracking is described.

5. CONCLUSIONS

If the nature of speech is mainly dynamic, then all phonemes (vowels and consonants) may be explained by a single theoretical framework and described by dynamic parameters defined in their time derivatives (deformation and rate of gestures, and corresponding direction and rate of trajectories in the acoustic space). According to the transition rate, fast transitions could produce consonants, slow transitions vowels, and middle rates produce diphthongs explaining the results reported in [27]. Then we might ask the real question: how to characterize vowels in typological terms?

6. REFERENCES

[1] Delattre, P. C., Liberman, A. M. and Cooper, F. S., 1955. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.* 27, 769-773.

[2] Öhman, S., 1966. Coarticulation in VCV utterances: spectrographic measurements. *J. Acoust. Soc. Am.* 39, 151-168.

[3] Lindblom, B., 1963. Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.* 35, 1773-1781.

[4] Kozhevnikov, V. A. and Chistovich, L. A. (1965) "Speech, articulation, and perception," JPRS-30543. NTIS, US Dept. of Commerce.

[5] Sussman, H. M., McCaffrey, H. A. and Matthews, S. A., 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. *J. Acoust. Soc. Am.* 90, 1309-1325.

[6] Stevens, K. N. and Blumstein, S., 1978. Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.* 64, 1358-1368.

[7] Lindblom, B. and Studdert-Kennedy, M., 1967. On the role of formant transitions in vowel perception. *J. Acoust. Soc. Am.* 42, 830-843.

[8] Nearey, T. and Assmann, P., 1986. Modeling the role of inherent spectral change in vowel identification. *J. Acoust. Soc. Am.* 80, 1297-1308.

[9] Strange, W., Jenkins, J. J. and Johnson, T. L., 1983. Dynamic specification of coarticulated vowel. *J. Acoust. Soc. Am.* 74, 695-705.

[10] Dorman, M. F., Studdert-Kennedy, M. and Raphael, L. J., 1977. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception and Psychophysics* 22, 109-122.

[11] Furui, S., 1986. Speaker-independant isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoust. Speech Signal Processing* 34, 52-59.

[12] Stevens, K., 1985. Evidence for the role of acoustic boundaries in the perception of speech sounds. In V. A. Fromkin, (Eds.), *Phonetic Linguistics. Essay in Honor of Peter Ladefoged*, Academic Press, Orlando, pp. 243-255.

[13] Fowler, C., 1980. Coarticulation and theories of extrinsic timing. *J. of Phonetics* 8, 113-133.

[14] Fant, G., 1960. *Acoustic theory of speech production*. Mouton, The Hague.

[15] Carré, R., Submitted. Dynamic properties of an acoustic tube: Prediction of vowel systems.

[16] Browman, C. and Goldstein, L., 1986. Towards an articulatory phonology. In C. Ewan and J. Anderson, (Eds.), *Phonology yearbook*, Cambridge University Press, Cambridge, pp. 219-252.

[17] Carré, R., Submitted. Production and perception of vowels without acoustic static targets.

[18] Chistovich, L. A., 1962. Temporal course of speech sound perception. In: *Proc. of the IV Int. Congr. Acoust.*, Copenhagen.

[19] Kent, R. D. and Moll, K. L., 1969. Vocal-tract characteristics of the stop cognates. *J. Acoust. Soc. Am.* 46, 1549-1555.

[20] Gay, T., 1978. Effect of speaking rate on vowel formant movements. *J. Acoust. Soc. Am.* 63, 223-230.

[21] Weismer, G. and Berry, J., 2003. Effects of speaking rate on second formant trajectories of selected vocalic nuclei. *J. Acoust. Soc. Am.* 113, 3362-3378.

[22] Adams, S. G., Weismer, G. and Kent, R. D., 1993. Speaking Rate and Speech Movement Velocity Profiles. *Journal of Speech and Hearing Research* 36, 41-54.

[23] Pitermann, M., 2000. Effect of speaking rate and contrastive stress on formant dynamics and vowel perception. *J. Acoust. Soc. Am.* 107, 3425-3437.

[24] Pollack, I., 1968. Detection of rate of change of auditory frequency. *J. Exp. Psychol.* 77, 535-541.

[25] Divenyi, P. L., 2005. Frequency change velocity detector: A bird or a red herring? In D. Pressnitzer, A. Cheveigné and S. McAdams, (Eds.), *Auditory Signal Processing: Physiology, Psychology and Models*, Springer-Verlag, New York, pp. 176-184.

[26] Chistovich, L. A., Lublinskaja, V. V., Malinikova, T. G., Ogorodnikova, E. A., Stoljarova, E. I. and Zhukov, S. J., 1982. Temporal processing of peripheral auditory patterns of speech. In R. Carlson and B. Grandström, (Eds.), *The representation of speech in the peripheral auditory system*, Elsevier Biomedical Press, Amsterdam, pp. 165-180.

[27] Miller, J. and Baer, T., 1983. Some effects of speaking rate on the production of /b/ and /w/. *J. Acoust. Soc. Am.* 73, 1751-1755.