3. REDEFINING THE SEGMENTATION PROBLEM

Within the context, defined above, speech-segmenting consists in researching an acoustic trajectory in the hope of tracking down targets, whether or not they are articulatorily met. As may be noticed, the larger problem of target identification can be made to pertain to acoustico-phonetic decoding, thanks to a grammar of distortions; as such a grammar of distortions can be inferred both from what is already known of co-articulation and from facts observed along the trajectory.

4. AN APPROACH THROUGH ANALYTICAL-MECHANICS

4.1. Usual Dimensions

Beside the already defined notions of velocity and acceleration, other dimensions can also be computed :

- curvature radius of the trajectory at point $M(t_n)$
- torsion of the trajectory at point $M(t_n)$

Whence it is possible to deal with the usual notions of rectilinear trajectory, stationary trajectory, etc. These notions can be extend over even longer temporal window-slits by associating, to each point $M(t_n)$, the variance-covariance matrix calculated over the m preceding points, using the m vectors $\{X_{n-m+1} \ldots X_n\}$. The first two proper directions (proper vectors) of this matrix can be assimilated to the directions of, respectively, the mean velocity vector $\underline{V}_n$ and the mean acceleration vector $\underline{G}_n$, on both of which the computations, aluded to above, can be run.

Now, if a mass is associated to point M, any directional alteration is the resultant of all forces applied to this point. It being assumed that clustering forces are frictionless, and that point M obeys strictly to the general laws of dynamics: point acceleration (whether positive or negative) is the resultant of attraction forces whose respective origins are the different targets --here considered as force fields.

4.2. Modelization

In order to extract interpretable path-portions from the trajectory, the following assumptions are made:

(a) the material point M moves towards one and only one target at a time,

(b) a target is considered met, whenever the trajectory becomes quasi-stationary,

(c) clustering forces are frictionless,

(d) the mean velocity $\underline{V}_n$ increases with speech output,

(e) a target is there but fails to be met, whenever the trajectory shows either a retrogression point or a sudden and marked directional change,

(f) around each target, there exists a force field the intensity of which decreases with speech output.

4.3. Experimentation

In an initial study, the p parameters of R^p to be retained were cues, otherwise used in speech analysis [Caelen et al, 81]. They are slow-variation cues, and thus the trajectories secured were sufficiently "smooth" to be meaningful. Over a preliminary corpus (isolated words pronounced by 10 speakers) the following observations were made: (Fig. 1)

(a) parameters are locally correlated according to phonemes; bringing out the existence of local clustering forces (or constraints). This should not come as a surprise, since we are dealing with co-articulation phenomena; but it allows (through intercorrelation-coefficient parameters) to quantify these phenomena.

(b) within a transitional phase between targets, the trajectory is quasi-linear (although this depends upon the coordinate system used).

(c) the trajectory is quasi-stationary whenever a target is met. A "Brownian movement" is then to be noticed around the target center.

(d) the trajectory does exhibit a directional alteration, if a target fails to be met.

(e) whenever speech-output rate becomes high, the number of such "failed" targets rises, while their mean reciprocal distances decrease.

(f) point M picks up speed as it leaves a target, and slows down as it nears the next one.

(g) there exists a grammar of distortions that makes it possible to superpose various speakers respective utterance trajectories.

4.4. Segmenting Automaton

On the basis of the preceding observations (a through g) it is possible, for the purpose of segmenting, to classify trajectories into three different types :

1 - "Brownian" trajectories (weak-amplitude motion about a target center) corresponding to a "target-met" detection procedure (TM).

2 - "Angular" trajectories (negative scalar product of mean velocities, retrogression point, slow down before odd point and speed up thereafter) corresponding to a "failed-target" detection procedure (FT). Note that the failed target always lies beyond the retrogression point.

3 - "Steady" trajectories (large curvature-radius, no odd point, maximum velocity reached about mid-course) corresponding to a transition-path detection procedure (T).

These three types of trajectory define the three different states assumed by an automaton whose transitional arcs are activated by TM, FT and T procedures.

5. CONCLUSION

The above makes it possible to look at segmenting, and subsequently at acoustico-phonetic decoding, from a new and maybe more advantageous angle : instead of researching discontinuity, we would resort to the formal instruments of mechanics (or data-analysis) to examine local variations in speech-trajectories that are represented in suitable spaces. Such a representation allows for an ascending description, from acoustics to phonology; while by-passing any a priori (even implicit) phonetic model. At the same time, it seems possible to find a grammar of distortions capable of superposing the several trajectories that correspond to one sequence uttered by several speakers. This kind of results, nevertheless, remains to be confirmed over large speech-corpuses and large numbers of speakers.

Acknowledgment: I wish to thank J.F. Malet (of California State University, Sacramento) for translating this text from French, and for suggesting various improvements as I wrote it.

BIBLIOGRAPHIC REFERENCES

[Abry et al, 85] C. Abry, C. Benoit, L.J. Boë and R. Sock, Un choix d'événements pour l'organisation temporelle du signal de parole, Proceedings GALF-CNRS XIVèmes JEP, Paris, 1985, pp.133-135

[Caelen et al, 81] J. Caelen and G. Caelen-Haumont, Indices et propriétés dans le projet ARIAL II, Proceedings GALF-CNRS, Processus d'encodage et de décodage phonétique, C. Abry, J. Caelen, J.S. Liénard, G. Perennou and M. Rossi eds., 1981, pp. 129-143

[Cohen, 81] J.R. Cohen, Segmenting speech using dynamic programming, JASA, Vol 69, n°5, 1981

[Fant, 73] G. Fant, Speech sounds and features, MIT Press, Cambridge, 1973

[Fujimura, 81] O. Fujimura, Temporal organization of articulatory movements as a multidimensional phrasal structure, Phonetica, Vol. 38, 1981, pp. 66-83

[Jakobson et al, 51] R. Jakobson, G. Fant and M. Halle, Preliminaries in speech analysis, MIT Press, Cambridge, 1951

[Ladefoged, 71] P. Ladefoged, Preliminaries to linguistic phonetics, The University of Chicago Press, 1971

[Lindblom, 83] B. Lindblom, Economy of speech gestures, in: P.F. MacNeilage ed., The production of speech, Springer-Verlag- Heidelberg, 1983

[Rossi, 83 M. Rossi, Niveaux de l'analyse phonétique: nature et structuration des indices et des traits, Speech. Com., Vol. 2, n° 2-3, 1983, pp. 91-106

[Zue, 83] V. Zue, The use of phonetic rules in automatic speech recogition, Speech. Com., Vol. 2, n° 2-3, 1983, pp. 181-186

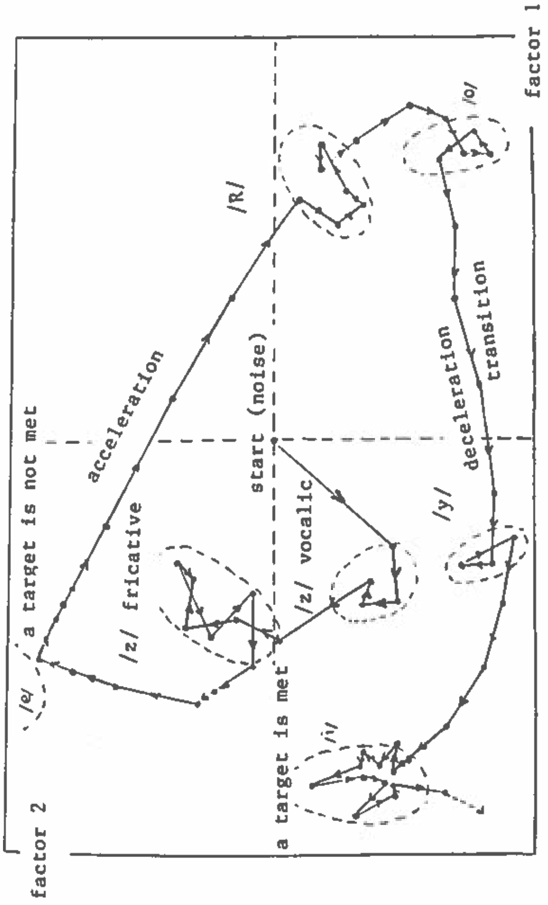

Fig. 1: Trajectory of the words "zéro-huit" /zeRo yi/