A SPECTRAL WARPING MODEL
A study of French Nasal vowels

P.F. MARTEAU, J. CAELEN & M.T. JANOT-GIORGETTI

Laboratoire de la Communication Parlée
I.C.P. Unité associée au CNRS (UA 368)
INPG/ENSERG - 46 avenue Félix Viallet - 38031 GRENOBLE CEDEX FRANCE

ABSTRACT

Most of the speech signal analysis methods deal with the spectral properties of homogeneous segments, assuming the properties are invariant on these segments. These methods do not take into account the dynamic aspects and the spectral warping between two contiguous segments. We propose a descriptive model which focuses on what does in fact vary in a set of successive spectra. This model is based on a time-domain statistic analysis of the frequency band energies of the spectra. We have experimented this approach by analysing the behavior of some French nasal vowels in a continuous speech corpus. We think that this sort of model is able to follow temporal transitions of different masses of energy. We show that these masses and the phases of their motion could be used as acoustic correlates.

INTRODUCTION

The speech signal is a time-domain varying signal. Whatever the kind of processing used, we always come back to a set of parameters which vary in time. (This set of parameters at the instant t represents the signal on a relatively short segment centered on t). If we suppose that the signal is represented by a m-parameters spectral vector (S), the evolution of this signal will be materialized by the motion of the point P representing the vector extremity ; with time, P follows a trajectory which reflects articulatory gestures in the R space. We can then notice on this trajectory relevant properties such as accumulation or turn-back points, rectilinear or curve segments (CAELEN 1986). If an accumulation point (a reached target) expresses a steady state easy to analyse, it is not the same for turn-back points (not reached targets) or transitions.

Some questions then arise : can we extract some laws relative to the spectral warpings in the vicinity of these transcient segments ? If these laws exist, can we interprete them as the trace in the acoustic space of the movement of the articulators.

From a perceptual level too, the role of the spectral change seems to be of great importance, even for the perception of vowels considered usually as monophtongs vowels. The modeling of these spectral changes are mainly based on formant transitions (LINDBLOM & STUDDERT-KENEDY 1967, GAY 1968, NEAREY & ASSMAN 1986, BROAD & CLERMONT 1987). The model that we suggest in order toanalyse acousticalevents is a mid term model of spectral warpings (its validity being reduced to a length of time about 100 ms).

We shall see that it is possible to extract information relative to the displacement of masses of energy which characterize the temporal signal variations. Finally, we shall evaluate this kind of model within the framework of a study on french nasal vowels in a continuous speech corpus.

1. THE MODEL

The acoustic signal is transposed in the frequency space by computingevery (t=5ms) a LPC spectrum (14 coefficients), defined on (m = 22) frequency bands, one Bark wide. The integration to one Bark brings us closer to the auditive transformation observed in man. Taking a temporal window W, correponding to the instants (1,...,q), we may note $J=\{1..q\}$ and $I=\{1..m\}$. On this window, the speech signal is thus represented by a rectangular array of real numbers: $\{S_i^j \mid i\in I;\ j\in J\}$.

We say that(S_i^j) is the energy of the i-th band for the j-th sample.

we may note $E=R^I$ the vectorial space of the real functions (f) defined by : $E=\{f \mid f^i\in R; i\in I\}$ and $G=R^J$ the vectorial space of the real functions (g) defined by : $G=\{g \mid g^j\in R; j\in J\}$. There are then two different ways to analyse the array(S_i^j) wether we consider the cloud C(I) of the functions S^i in the space G or the cloud C(J) of the functions S^j in the space E.

C(J) can be interpreted as a time-domain trajectory in the space E and C(I) as a frequency-domain trajectory in the space G. We find again the time/frequency duality. The two clouds can be then analyzed in terms of inertia, in order to extract the main tendency of the temporal or frequency evolutions. This is the purpose of the **factorial analysis (BENZECRI 1973).** This type of analysis applied to speech processing is not new (CARTIER & GRAILLOT 1974, CAELEN & VIGOUROUX 1983). We insist on the fact that here the objects are not taken from a set of independant observations : the objects are linked by a time relation. The theory of the factorial analysis tells us that the two previous points of view are equivalent (or dual). This means that the inertia characteristics of the two clouds are linked by a bijective relationship. **This is why, superposing the two spaces E and G, we may interpret simultaneously the transitions of the signal as the dislocation of some masses of energy in the time domain and in the frequency domain.**

We consider then the cloud C(J) in the space E. We may note $\{\nu_j \mid j\in J\}$ a set of positive numbers : ν_j is the weights assigned to the j-th sample (spectrum). (In practice we use the Hamming window or the rectangular window). Let's note $\bar{w}$ the sum of the weights. If we fit the space E with the euclidian distance (d), (E,d) is a topologic space. The distance between two objects j and j' is given by :

$d(j,j')^2 = \sum\{(S_i^j - S_i^{j'}); \; i \in I\}$

We call G the gravity center of the functions S : $G_i = \sum\{(w_j/\bar{w}) S_i^j ; | \; j \in J\}$ and $\{\sigma_{ii'} | i \in I; \; i' \in I\}$ the symetric square array representing the quadratic inertia function of the cloud C(J).

$\sigma_{ii'} = \sum\{ w_j S_i^j S_{i'}^j \mid j \in J\} - G_i G_{i'} \bar{w}$.

U = $\{u^i | i \in I\}$, the set of the orthonormal eigen vectors for the application σ ; $\sigma(u)^i = \lambda u^i$

V = $\{v^j | j \in J$; u U G , the corresponding main components $v^j = \sum\{u^i(S_i^j - G_i) | i \in I\}$.

Then, if we consider the sub-system U(p) of the p-first eigen vectors, assigned to the p-highest eigen values , we obtain a p-order estimate of the array S which is given by :

$$\hat{S}_i^j = G_i + \sum\{(\sqrt{\lambda(h)} v^j(h) u_i(h)); \quad h=1..p\} \quad (1).$$

S is the p-order array which minimize the criterion

$$\sum\{(\hat{S}_i^j - S_i^j)^2 \mid i \in I; j \in J\}$$

What is the meaning of the model described by the equation (1) ? For the speech signal analyzed on a temporal window W having an order of magnitude of 100 ms, the model is reliable when n equals 2 or 3. This means that on this analysis window, the trajectory of the representative spectral evolution's point is approximatively contained in a 2 or 3 dimensional sub-space. The (u(h)) forms can be interpreted as frequency mask functions, balanced by the ($\sqrt{\lambda(h)}$v(h)) coefficients. Thus, we can compare them to the spectral cues defined by ROSSI or CAELEN (ROSSI & al 1983, CAELEN & CAELEN-HAUMONT 1981). Such a model is entirely related to the choice of the analysis window. Nevertheless, we can test its reliability to small time translations, i.e. its predictive power. Let the temporal analysis window be W = $\{1..q\}$ and $\{u(h), v(h), \lambda(h); \; h=1..p.\}$ the p-order estimate model. Let the spectrum S^{q+1} calculated at instant (q+1). Its projection in the sub-space E gives for this spectrum the values of the function: $\sqrt{\lambda(h)} v^{q+1}(h)$ The euclidian distance between the spectrum and its reconstruction through the model is a rupture criterion, which is expressed by

$$\delta^{q+1} = \sum\{(S_i^{q+1} - \hat{S}_i^{q+1})^2 \; i \in I\} \quad (2)$$

assuming that the window W includes a part of a steady state, δ^{q+1} will stay small and the same will happen with δ^{q+i} and i $\geqslant$ 1. The model will be reliable with regard to a small future insight. Nevertheless, for a larger temporal insight, we expect the criterion quickly to deteriorate.

This means that the model is not reliable enough, that the past is not sufficient to explain the future. Hence, the analysis window parameters : position and length of time, are of great importance when considering the convergence of the model and its power for describing spectral warpings with coherence. The choice of the appropriate window is a complex problem. In our opinion, this choice may be conditioned by the search for instabilities, since in the way it is built the model focuses on these instabilities.

the rupture criterion exposed before (relation N°2) can produce an automatic segmentation of the speech signal.

The results of such a segmentation will be presented later.

exposed before (relation N° 2) can then produce an automatic seg- mentation of the speech signal. The results of such a segmentation will be presented later. Let us say simply that a signal fragment may be split up into smaller fragments by finer and finer downward analysis. This is performed by eliminating the successive instabilities detected on topologic criteria. We thus realize a tree of sub-models representing the speech signal on a given time interval. The leaves of this arborescence are close to homogeneous phones (CAELEN 1981) and describe acoustical events.

Following this, we try to show that in a suitable window, i.e. a window with few instabilities, the basis $\{u(h) | h=1..p\}$ and the evolution of their associated components $\{\sqrt{\lambda} v(h) | h=1..p\}$ express energy displacement interpreted as motion towards energy targets. We will notice that these masses of energy have not necessary relationship with formants. We have chosen for this study the French nasal vowels /ɑ̃/, /ɔ̃/, /ɛ̃/, /œ̃/) because of their complex structures which have given rise to many studies. We will then try to classify different observation cases, in order to extract laws and descriptive rules for the phonetic feature of nasality /+ NASAL/.

2. THE NASAL VOWELS

From an articulatory level, the nasal vowels are the result of an acoustic coupling between oral and nasal cavities. The coupling occurs approximatively at the middle of the vocal cavity, between the lips and the glottis. The consequence of this coupling is the displacement of the first natural oral formant and the appearance of a pole-zero pair in the transfer function (FUJIMURA & LINDQUIST 1971, MRAYATI 1976, MAEDA 1982, HAWKINS & STEVENS 1985...). According to the authors, the resulting acoustical correlates generally affect the low frequency part of the spectrum ; these correlates are in fact the frequency displacement of the first formant and the widening of the bandwith. A formant with high damping also appears at the viciniy of 300 Hz. Some studies try to introduce more accurate dynamic information.

MERMELSTEIN set up four acoustical parameters which are relative variations of the energy in the frequency bands [0-1], [1-2], [2-5 kHz] and the centroïd of the frequency band [0 - 500 Hz].

FENG proposed the nasopharyngal target concept which is characterized by the appearance of two formants : one is located near 300 Hz, the other located near 1000 Hz. This concept arises from articulatory simulation (FENG & al 1985).

CHENG showed from psychoacoustical data the possibility of a balance between two masses of energy centered in the vicinity of 300 and 1000 Hz (DELATTRE 1969, CHENG 1987).

Hence, nasality should be characterized by spectral deformations from an initial spectrum corresponding to an oral or quasi-oral vowel, towards a final spectrum with greater energy in the neighbourhood of 300 and 1000 Hz. Since we observed a sort of analogy between the articulatory and perceptive concepts, can we encounter similar properties in the acoustical space ? We shall see that the spectral warping model brings some significant elements.

3. THE STUDY

3.1. corpus

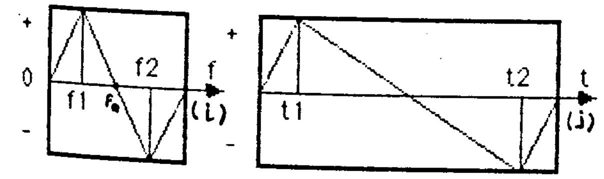
The corpus includes a 45 words text, the average time duration of which is 29 s. The text is an abstract of a paper published in **"Science et vie"**. The records of ten speakers (5 women and 5 men) are used. These speakers were requested to speak naturally and comprehensibly. The records were made in a quiet room. A N 4420 Radiola recorder was used.

"D'éminents biologistes et d'éminents zoologistes américains ont créé pour des vers géants, un nouveau phyllum dans l'actuelle classification des nombreuses espèces vivantes. Ces longs vers prospèrent sur le plancher marin des zones sous-marines profondes. Des sources thermales chaudes y maintiennent une température moyenne élevée."

The corpus was manually labelled by an expert in phonetics, according to labelling principles based upon a spectral analysis (VIGOUROUX & CAELEN 1985). Selection of the nasal vowels is then performed automatically from a label file. The corpus includes 160 nasal vowels (7/ɑ̃/ 5/ɔ̃/, 4/ɛ̃/) for ten different speakers. In this study, we do not take into account the differences between and

3.2. Interpretation procedure

The interpretation is carried out through qualitative analysis of the pair$(u(h), \sqrt{\lambda(h)} v(h) | h=1..p)$. For instance let the pair $(u, \sqrt{\lambda} v)$ be expressed as follow :



from t_1 to t_2, energy decreases in the vicinity of frequency f1 and increases in the vicinity of frequency f2. Energy in the vicinity of f_0 is constant for this transition. The extreme values of the ($\sqrt{\lambda(h)}$v(h)) functions (for t1 and t2) can be interpreted as targets.

- t1 is the target characterized by spectra the energy of which is maximal in the vicinity of f1 and minimal in the vicinity of f2.
- on the contrary, t2 is the target featured by spectra the energy of which is minimal in the vicinity of f1 and maximal in the vicinity of f2.

Thus, analysis of the segment /a/ (Fig. 1.1 and 1.2) shows that two targets exist :

- The first one corresponds to the energy masses centered around the following frequencies (450-540 Hz/1550-1790 Hz). These two masses are connected to the two first formants of the /a/. The stability of this target implies an accumulation point on the trajectory.
- The second one corresponds to the energy masses centered around the following frequencies (450-540 Hz/2060-2350 Hz) and deals with the increase of energy in the vicinity of the first formant and a dislocation of the second formant towards higher frequencies (influence of /ki/ context). This second target, not reached, creates a turn back point on the trajectory.

The analysis of segment /ɑ̃/ (fig. 2.1 and 2.2.) shows as well two instable targets.

- One near the frequencies (540-650 Hz, 1130-1340 Hz) corresponds to an oral vowel.
- The other near the frequencies (260-320 Hz, 880-1000 Hz) corresponds to a nasal vowel:

These two targets are represented by two turn-back points on the trajectory on either side of u axis. The trajectory is nearly rectilinear between the two targets.

3.3. Results

For the whole corpus results are plotted on figures similar to figures (1.2) and (2.2).

The interpretation is manual.

- We may note by (CN) the presence of a target characterized by a simultaneous increase of energy in the frequency bands [200-500 Hz] and [760-1340 Hz]
- We may note by (CH) the presence of a target characterized by an increase of energy in the frequency band [760-1340].
- We may note by (CB) the target characterized by an increase of energy in the frequency band 200-500 Hz
- We may note by (ER) the presence of an error, when no interpretation is possible.

The results are reported in fig. 3 at the end of this paper.

On the whole nasal vowels of the corpus, all context taking into account, the following occurences appear : CN = 65 %, CH = 29,5 %, CB = 0 %, ERR = 5,5 %.

Based on this data, the energy increase in the band [760-1340 Hz] appears to be a significant acoustical correlate since it is present in 94,5 % of the cases. In 35 % of the cases, the model is not sensitive to any significant variation in the band [200-450 Hz]. This means that the low frequency formant variation is not always necessary to perform a French nasal vowel. Analysis-synthesis techniques, using a formant synthesizer, leads to similar conclusions for nasal vowels of French spoken in Montreal (LAFFERIERE & O'SHAUGHNESSY 1986).

Nevertheless it is interesting to observe that results obtained from a dynamic articulatory model (FENG 1986) and pyschoacoustical results featuring the sensitivity of controlled parameters of synthesizer with respect to the perception associated to nasality (FENG 1987), are partially consistent to the results given by the model of spectral deformations : the existence of two dynamic energy masses (64% of the cases). The first one near 300 Hz, the other one, dominant, near 1000 Hz. The target concept is represented by the evolution of the ($\sqrt{\lambda} v$) function and the notion of energetic balance could be interpreted by the structures of the (u) forms.

CONCLUSION

We have proposed a model which takes into account local transitions of energy masses. These masses are computed. Thus we do not have to face the problem of detection. These energy masses are usuallyassociated with formants when these are dynamic either by frequency translations or by widening the bandwidth.

But they also take into account phenomena such as spectral flattening. This spectral warping model has allowed us to relocate roughly the results of articulatory and psychoacoustic analysis within the framework of the study of the French nasal vowels.

We think that the procedure of interpretation may be automated by the use of expert system. The experience and the knowledge acquired from spectogram reading systems will be extremely usefull. Finally the use of such a model is not limited to the field of acoustic-phonetic decoding. In our opinion it could also be used in low rate coding.

BIBLIOGRAPHY

BENZECRI J.P. (1973)
L'analyse des données
Dunod (1973).

BROAD D.J., CLERMONT F. (1987)
A methodology for modeling vowel formant contours in CVC context
J. Acoust. Soc. Am., 81 (1), 155-665.

CAELEN J., CAELEN-HAUMONT G. (1981)
Indices et propriétés dans le projet ARIAL II
Proceedings GALF-CNRS "Processus d'encodage et de décodage phonétique".
C.Abry, J.Caelen, J.S.Lienard, G.Perennou & M.Rossi.

CAELEN G., VIGOUROUX N. (1983)
Les indices de distribution spectrale. Etude comparative au travers de 2 analyses discriminantes monolocuteur et interlocuteur
Speech Communication, 2, 133-136.

CAELEN J., VIGOUROUX N., (1985)
Une base acoustique et phonétique hiérarchisée : des faits aux connaissances
Actes du Symposium Franco-Suédois sur la parole, GRENOBLE.

CHENG Y.M., GUERIN B. (1987)
Nasal vowel study : formant structure, perceptual evaluation and neural representation in a model of the peripheral auditory system
Institut de la Communication Parlée, Bulletin n° 0.

DELATTRE P. (1969)
The General Phonetic Charactéristic : Final report
US Dept. of Health, Education & Welfare. Office of Education Institute of International Studie.

FENG G., ABRY C., GUERIN B. (1985)
How to cope with nasal vowels ? Some acoustic boundary poles
Actes du Symposium Franco-Suédois sur la parole, GRENOBLE.

FUJIMURA O., LINDQUIST J. (1971)
Sweep-tone measurement of vocal-tract characteristics
J. Acoust. Soc. Am., 49 (2), 541-558.

GAY T. (1968)
Effects of speaking rate on diphthong formant movements
J. Acoust. Soc. Am., 44, 1570-1573.

GAY T. (1978)
Effects of speaking rate on vowel formant movements
J. Acoust. Soc. Am., 63, 223-230.

HAWKINS S., STEVENS K.N. (1985)
Acoustic and perceptual correlates of Non-Nasal/Nasal Distinction for vowels
J. Acoust. Soc. Am., 77, 1560-1575.

LAFERRIERE F., O'SHAUGHNESSY D. (1986)
Analyse-synthèse et études de règles acoustiques de production avec un synthétiseur à formant
15ème J.E.P., 11-14, AIX EN PROVENCE.

LINDBLOM B., STUDDERT-KENNEDY M. (1967)
On the role of formant transitions in vowel recognition
J. Acoust. Soc. Am., 42, 830-843.

MAEDA S. (1984)
Une paire de PICS spectraux comme corrélat acoustique de la nasalisation des voyelles
13ème J.E.P., 223-224, BRUXELLES.

MERMELSTEIN P. (1977)
On detecting nasals in continuous speech
J. Acoust. Soc. Am., 16 (2), 581-587.

MRAYATI M. (1976)
Contribution aux études sur la production de la parole
Thèse Doct. d'Etat, I.N.P. GRENOBLE.

ROSSI M., NISHINUMA Y., MERCIER G. (1983)
Indices acoustiques multilocuteurs et indépendants du contexte pour la reconnaissance automatique de la parole
Speech Communication, 2, 215-217.

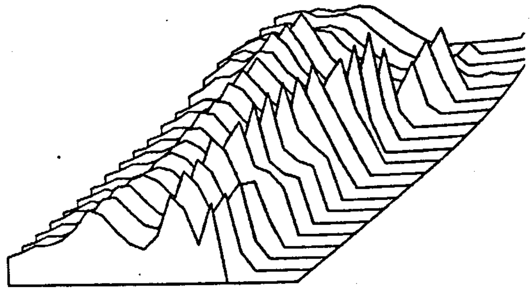


FIG N° 1.1 : The set of successive spectra S_i^t on the /a/ segment.

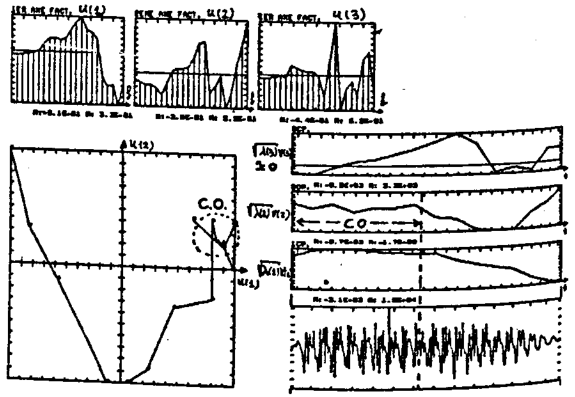


FIG N° 1.2 : The model on the /a/ segment; The eigen vectors (u) and the associated main components (Gv)
The trajectory in the space (u(1),u(2))

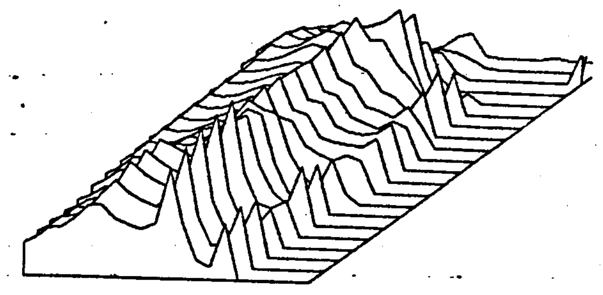


FIG N° 2.1 : The set of successive spectra S_i^t on the /ã/ segment.

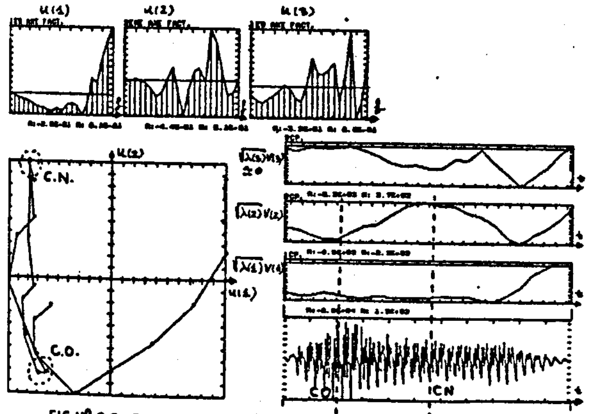


FIG N° 2.2 : The model on the /ã/ segment, The eigen vectors (u) and the associated main components (Gv)
The trajectory in the space (u(1),u(2))

THE RESULTS

Vowel	Context	CN %	CH %	CB %
/ɑ̃/	/mɑ̃b/	70	30	0
	/mɑ̃z/	50	50	0
	/eɑ̃m/	90	10	0
	/dɑ̃l/	60	30	0
	/vɑ̃t/	60	30	0
	/ʃɑ̃s/	80	10	0
	/tɑ̃p/	80	10	0
/ɔ̃/	/mɔ̃k/	30	70	0
	/jɔ̃d/	60	30	0.
	/nɔ̃b/	80	10	0
	/lɔ̃v/	60	40	0
	/tɔ̃d/	80	20	0
/ɛ̃/	/kɛ̃ /	80	20	0
	/rɛ̃d/	60	30	0
	/ ɛ̃n/	60	30	0
	/mɛ̃t/	50	40	0

FIG. No 3