LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI

De la voix humaine considérée comme un instrument de musique

Les cahiers de l'analyse des données, tome 8, n° 2 (1983), p. 181-186

<http://www.numdam.org/item?id=CAD_1983__8_2_181_0>

Les Cahiers de l'Analyse des Données
Vol. VIII - 1983 n° 2 - p. 181 186

DE LA VOIX HUMAINE CONSIDÉRÉE COMME UN INSTRUMENT DE MUSIQUE

[VOIX]

par J.P. Benzécri (1)

Ich gebe zu, dasz mein Ausweg
veilleicht von vornherein wenig
wahrscheinlich erscheinen mag,...
Aber nur wer wagt, gewinnt...

W. Pauli *

Il est bien connu qu'il y a trois qualités physiologiques d'un son : l'intensité, la hauteur et le timbre ; il est facile de produire des sons sinusoïdaux d'intensité et de hauteur quelconque ; en revanche l'analyse et la synthèse du timbre sont beaucoup plus complexes. Voici un siècle déjà, que Helmoltz réalisait la synthèse des timbres de voyelles par un système de diapason donnant les harmoniques d'un fondamental ; le vocodeur a repris analyse et synthèse avec les moyens électroniques du milieu du XX-ème siècle. Il ne fait pas de doute pour nous toutefois que seul le traitement digital est assez précis pour achever l'étude de la parole.

Le modèle de la voix humaine, fortement suggéré sinon totalement confirmé après la thèse de T. Moussa, est celui d'un instrument de musique au sens de M. Matthews.

L'instrument ou plutôt la partition qu'il exécute, est décrit comme il sied par trois fonctions du temps : une intensité, une fréquence fondamentale et un timbre : toutefois alors que l'intensité et la fréquence sont des fonctions usuelles à valeur réelle, le timbre est à chaque instant une fonction de la fréquence (fonction de R). En fait cette fonction peut être caractérisée et reconstruite (avec la précision utile) à partir de facteurs (au sens de l'a. des corr.) dont le nombre ne dépasse pas 7.

Voici donc la formule fondamentale de l'amplitude :

$$x(t)=I(t)\ \Sigma\{Env(n\omega(t);t)\ \sin(n\omega(t)t+\varphi(n))\mid n=1,\ldots\}\quad ;$$

ici : $I(t)$ désigne l'intensité instantanée (au temps t) ;

$\omega(t)$ est la fréquence instantanée (au temps t) ;

$Env(\omega;t)$ est la fonction d'enveloppe spectrale ou de timbre.

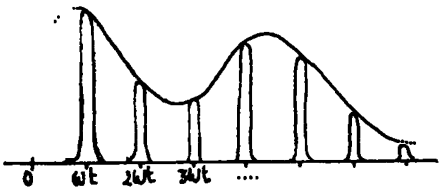
---

* *in Collected Sc. Papers Vol II p. 1317 ; Interscience 1964 In* Aufsätze und Vorträge über Physik und Erkenntnistheorie W. F. Weisskopf ed. Vieweg & Sohn, Braunschweig, 1961 p. 160.

(1) *Professeur de statistique. Université P. et M. Curie.*

La sommation est faite sur la suite indicée par n des harmoniques du fondamental ; les phases $\varphi(n)$ des harmoniques successifs n'ont *a priori* aucune importance : il faut seulement maintenir la cohérence temporelle des phases de chaque harmonique.

Un son ainsi défini peut être schématisé par une suite de pics, le fondamental et ses harmoniques, logés sous une enveloppe !



Eventuellement, on remplacera cette suite de pics par un bruit coloré pour obtenir la voix chuchotée (sans tonalité musicale).

Le modèle ainsi proposé est universel : il convient à toutes les langues : évidemment la fonction d'intonation $\omega(t)$ et surtout l'enveloppe $Env(\omega;t)$ sont propres à chaque langue. On distinguera donc d'une part le synthétiseur universel, capable de réaliser physiquement la formule fondamentale à partir de la donnée de $I(t)$, $\omega(t)$, $Env(\omega;t)$ et d'autre part l'étude analytique qui fournit les paramètres, ou si l'on peut dire cette partition.

La première partie du programme d'analyses est dans la thèse de T. Moussa : elle consiste à calculer $I(t)$ puis $Env(\omega;t)$ en fonction d'un signal échantillonné à 20Kh : comme expliqué dans [*PAROLE I*] (G. Charbonneau et T. Moussa : C.A.D. Vol VI n° 2 ; 1981) on considère des tranches successives de parole avec une fenêtre glissante, d'où des spectres instantanés (calculés par transformation de Fourier) ; et pour chaque spectre on calcule un fondamental et une enveloppe - laquelle est complètement décrite par la formule de reconstitution usuelle de l'a. des corr. en fonction de 7 facteurs qui deviennent donc les 7 dimensions du timbre. Dès lors il y a neuf dimensions instanées $I(t)$, $\omega(t)$ et les sept facteurs de $F_1(t)$ à $F_7(t)$.

Le choix de la largeur de la fenêtre temporelle doit répondre à des exigences contradictoires : être assez large pour offrir une détermination précise de $\omega(t)$; être assez étroite pour fournir véritablement une enveloppe instantanée et non un timbre moyen couvrant des changement profonds : en fait il conviendra dans l'avenir d'utiliser deux largeurs, l'une pour le calcul de $\omega(t)$ l'autre pour les $F_n(t)$.

La deuxième étape de l'analyse est le découpage de la chaîne parlée en phonèmes, plus exactement, distinction essentielle, en phones. De notre point de vue, il s'agit, en bref de discrétiser la variable de timbre ; c'est là un problème sur lequel nous reviendrons. Mais auparavant, il convient de disposer d'un programme de synthèse universel, afin de vérifier que le codage à 9 dimensions permet une reproduction fidèle de la parole.

Le premier essai sera de synthétiser un timbre constant quelconque, parmi ceux obtenus par l'analyse de la parole : cela revient à donner aux $F_n(t)$, donc à Env une valeur indépendante de t. Ce que nous affirmons d'après les résultats de la thèse déjà citée, c'est qu'on obtiendra ainsi sous forme stationnaire, non seulement

des timbres de voyelles, mais des timbres de consonnes et même de consonnes occlusives ; le cas des occlusives sourdes (t,k) pouvant toutefois réserver des surprises (leur réalisation requerrait du bruit coloré).

Revenons à l'analyse : il s'agit d'étiqueter des segments successifs de la chaîne parlée. T. Moussa a obtenu sur la segmentation des résultats assez bons. Mais le fond du problème est que ni le découpage ni l'étiquetage ne peuvent correspondre à la transcription classique du discours en une suite de phonèmes (pour ne rien dire des lettres généralement peu fidèles à la prononciation réele). Que le f soit une entité du système de la langue française (ou d'une autre langue...) est une chose : qu'il soit prononcé v, ou p, ou p-h ou disparaisse en est une autre;seul nous intéresse, pour l'instant, ce qui est prononcé : la rectification des sons et la reconnaissance des mots ne peuvent venir qu'ensuite : prétendre reconnaître sans analyse des mots ou seulement des phonèmes peut conduire à des résultats immédiats non nuls ; mais pour donner à la reconnaissance de la parole toute la perfection dont elle est susceptible, il faut une analyse phonétique parfaite. Selon la terminologie des linguistes, de même qu'on appelle phonèmes les unités du projet sonore entre lesquelles existent des différences pertinentes pour le sens, il convient d'appeler phones les unités de la réalisation sonore. Cette partition de l'ensemble infini des timbres produits, en un ensemble fini de phones donne à réfléchir ! Pourquoi faire une partition plutôt que d'accepter un continuum ? La partition est-elle universelle ? propre à chaque langue ? ou même à chaque sujet parlant ?

Selon nous, il y a vraisemblablement dans le continuum des timbres un domaine propre à chaque langue, avec une partition de ce domaine en quelques dizaines de phones : qui sont des maxima locaux de densité d'utilisation plutôt que de pics isolés ; les excursions en dehors du domaine se produisant selon les fantaisies individuelles et aussi les transitions entre phonèmes. L'infinie variété des voix propres aux diverses personnes s'expliquent moins par des différences de timbres que par des différences dans l'intonation, c'est-à-dire dans la variation de $\omega(t)$ qui accompagne celle du timbre.

Au fond ce problème de la partition a une importance scientifique essentielle pour la linguistique ; mais techniquement on est assuré *a priori* qu'il pourra toujours être résolu par une classification automatique qui donne un système fini de classes aux centres desquelles on puisse rattacher tout timbre nouveau ; que ces classes soient nettement séparées ou contiguës entre elles ! Dès maintenant partant d'un discours transcrit en une suite de timbres décrits à tout instant par un point $(F1(t),\ldots, F7(t))$ dans l'espace à 7 dimensions, avec comme a pu le faire T. Moussa un étiquetage de chaque timbre par un phonème ou une transition interphonémique, il s'impose de faire une C.A.H. ; d'abord sur les timbres étiquetés par des phonèmes purs, puis sur tous les timbres recueillis afin de vérifier que les classes obtenues s'interprètent bien phonétiquement ; et surtout qu'en assimilant chaque timbre au timbre moyen de sa classe, on a un codage discret du timbre qui permet une synthèse satisfaisante ; la reconstitution d'un discours qui pour l'oreille équivaut au discours initial.

Techniquement le matériel requis pour de telles recherches, comprend un convertisseur de tension en chiffres et chiffres en tension ; un ordinateur de capacité modeste ; et afin d'opérer en temps réel, une transformée de Fourier spécialement programmée autour d'un processeur propre. Quant à l'informatique, il faut, outre les programmes classiques d'analyse des données, les programmes de

reconnaissance de la hauteur (ou ton) $\omega(t)$, et de dessin de l'enveloppe spectrale en vue de son analyse factorielle : cette étape a été franchie par T. Moussa. Reste à écrire le programme de synthèse, piloté par la données des 9 fonctions $I(t)$, $\omega(t)$, $Fn(t)$: de tels programmes sont classiquement utilisés par les musiciens depuis les travaux de M. Matthews et coll. à la compagnie Bell ; en France J. C. Risset et G. Charbonneau maîtrisent ces programmes.

L'essentiel est d'expérimenter avec ces instruments pour analyser et synthétiser des voix de toutes langues parlées chantées ou chuchotées par des personnes de tout âge et de tout sexe !

L'enjeu scientifique la connaissance de la voix, suffirait à animer des recherches obstinées. Les applications techniques rendent ces recherches urgentes : la compression du signal parlé et surtout la commande des automates par la voix sont aujourd'hui l'objet de travaux coûteux, acharnés mais peu féconds car la portée en est limitée par l'absence de connaissances fondamentales.

La résistance des électroniciens de l'acoustique à un traitement complètement digital, l'opposition des tenants de techniques de reconnaissances spécialisées telles que le codage prédictif, aux méthodes universelles de l'analyse des données, sont cause que la voie proposée ici et déjà explorée par T. Moussa reste largement ouverte à tout chercheur disposant du matériel modeste que nous avons dit.

Audaces Fortuna Juvat !

NOTE . *Sciences et techniques dans la synthèse de la parole.*

Du point de vue de la communication entre homme et machine, faire la synthèse de la parole, c'est permettre à la machine d'être comprise de l'homme : c'est en un certain sens, un problème déjà résolu; tandis que le problème inverse de l'analyse (faire comprendre l'homme par la machine) n'a reçu que des solutions partielles (pour quelques dizaines de mots par une seule voix, par exemple) et nullement satisfaisantes (les taux d'erreur étant toujours importants). Mais au fond, la synthèse ne peut être séparée de l'analyse ; d'autre part celle là est au confluent de plusieurs sciences, et les solutions techniques actuelles qui visent à tout résoudre en une seule étape sont selon nous limitées à la fois quant à la simplicité et quant à la qualité.

Pour la communication d'homme à homme, les deux problèmes de l'analyse et de la synthèse sont déjà en partie posés, et résolus, depuis l'invention de l'écriture. Nous ne tenterons pas ici de faire l'histoire longue et complexe des diverses formes d'écriture , mais il importe de voir quels obstacles les imperfections de l'écriture opposent à la synthèse automatique de la parole ; et aussi dans quelle mesure tout traitement automatique (qu'il soit une analyse ou une synthèse) comporte en lui-même une conception implicite, plus ou moins parfaite, de l'écriture.

Voici d'abord une conception naïve de la synthèse qui présuppose la perfection de l'écriture : il n'y a qu'à émettre successivement toutes les lettres du texte à synthétiser. Quiconque à buté sur la lecture d'un mot étranger, sans même en trouver la prononciation sur une grammaire ou un dictionnaire, connaît les limites de cette conception ! Au reste les difficultés varient grandement de langue à langue : un texte espagnol imprimé usuel n'offre strictement aucune difficulté ; il en est de même pour un texte arabe

complètement vocalisé ; au contraire un texte arabe usuel ne peut être lu correctement que par celui qui le comprend, une forme telle que كتبت (prétérit du verbe écrire) admettant par exemple 8 lectures différentes selon qu'on la comprend à l'actif ou au passif ; et aux personnes 1, 2M, 2F ou 3F (M = masculin ; F = féminin). L'orthographe particulièrement difficile du français, ou la phonétique capricieuse de l'anglais offrant d'autres difficultés ; sans oublier la place de l'accent qui n'obéit à des règles strictes que dans certaines langues...

Interrogé en 1950, un linguiste aurait répondu qu'il ne s'agissait en bref que de donner du texte à lire une transcription phonémique, version scientifique parfaite des prononciations figurées en usage depuis longtemps ; transcription qui est à peu près celle qu'offre l'écriture espagnole ; ou l'écriture arabe complète. Et dès le début du traitement des textes sur ordinateur, on vît naître des programmes visant par exemple, à transformer un texte français, saisi dans son orthographe usuelle, en une suite de phonèmes. Ces programmes sont de qualité diverse ; mais assurément aucun n'est parfait; car même si les difficultés sont moindres que pour l'arabe non vocalisé, la transcription phonémique sans faille du français requiert une véritable compréhension du texte, laquelle n'apparaît pas automatisable en l'état présent de la sémantique.

D'ailleurs cette conception phonétique stricte laisse dans l'ombre deux très grandes difficultés : l'accentuation, et la réalisation des phonèmes enchaînés.

En toute rigueur, si l'accentuation des mots isolés est assez bien étudiée, l'intonation des phrases qui en est inséparable, est mal connue : et les meilleurs spécialistes ne semblent pas avoir de système parfait pour décrire et donc noter avec clarté et concision cette sorte de ponctuation orale que constitue l'intonation.

Quant à la réalisation des phonèmes enchaînés, l'école phonologique de Prague (e.g. le Prince N. Troubetskoy ; ou R. Jakobson...) a marqué la distinction entre la suite des phonèmes, (unités distinctives), où en bref toute substitution est susceptible, dans un contexte convenable, de produire un changement de sens (e.g. p ≠ v parce que pin ≠ vin) ; et la diversité des réalisations d'un même phonème selon le contexte (e.g. en arabe il n'y a ni phonème v ni phonème p ; mais un "b" pourrait se prononcer " p " ; ou un " f ", comme "v"...). Quel que soit l'intérêt de cette distinction péremptoire, elle laisse à faire après l'établissement supposé parfait du système phonémique d'une langue donnée, un travail de phonétique qui, croyons nous n'a jamais été achevé pour aucune langue sur une base expérimentale assez large.

En réalité, avec l'intonation et la réalisation des phonèmes enchaînés, s'introduisent toutes les variations individuelles de la parole ; lesquelles selon nous, ne peuvent être décrites adéquatement qu'après avoir achevé l'analyse de la parole ; principalement par un détecteur digital de mélodie ; et la détermination continue de facteurs de forme de l'enveloppe spectrale. Le langage de commande pour une synthèse parfaite, n'est autre que le langage formel qui résulte d'une analyse achevée. Sans celle-ci nul partage entre la langue système propre à une communauté qui échange des discours , et les latitudes de réalisation individuelles, où la synthèse doit faire choix d'une règle neutre !

Ainsi un appareil unique qui transforme un texte français saisi dans son orthographe usuelle en un message sonore doit résoudre, vaille que vaille, des problèmes très divers qui relèvent de disciplines différentes ; et, pour certains desquels est inconnue la forme-même que doit revêtir la solution...

Un problème clair, mais dont la solution requiert analyse morphologique et consultation de dictionnaire : dans *perdent*, *"ent"* = e; dans *présent "ent"* = an*; avec pour quelques cas, tels que la prononciation de *plus* en plu ou pluss, des perspectives sémantiques inquiétantes (j'en veux "pluss" ; mais : je n'en veux "plu").

Des problèmes obscurs : l'intonation ; les variations de réalisation des phonèmes avec le contexte... : ici le traitement des données passe par une analyse de la chaîne parlée ; aboutissant, nous l'avons dit à une nouvelle forme de description.

Les solutions techniques actuelles, qui dans la synthèse assemblent des syllabes ou des couples de phonèmes et parfois des enregistrements de mots entiers, plutôt que des suites d'éléments, sont à la solution idéale, ce que les premières écritures idéogrammatiques et syllabiques apparues il y a 5.000 ans, étaient à la notation phonémique (pour ne rien dire d'une notation encore inconnue de la réalisation du discours).

Nous conclurons qu'en bonne science l'analyse ne se sépare pas de la synthèse : car sans analyse achevée, nulle synthèse parfaite ; sans l'épreuve d'une synthèse fondée sur elle, nulle analyse validée.

---

* *Une ambiguïté frappante est celle entre "couvent" nom, et "couvent" verbe couver, 3ème personne du pluriel du présent de l'indicatif.*