Extrait du livre

"Acoustique - Informatique - MusiquE"

publié aux Presses des Mines par
B. d'Andréa-Novel, B. Fabre et P. Jouvelot
(ISBN 978-2911256-60-8)

Chapitre 11 (version d'auteur - Thomas Hélie)

Table des matières

Chapitre 11

Vocoder par LPC

Thomas Hélie

Peut-on faire "parler" la musique ou, plus généralement, un son ?

Cette question a passionné musiciens, luthiers, ingénieurs et chercheurs qui ont apporté régulièrement des solutions avec les techniques de leur époque. Ce chapitre est destiné à la réalisation de travaux pratiques et l'étude d'une méthode utilisée en musique mais aussi dans les codeurs de voix de téléphones mobiles : le "vocoder par LPC" (Linear Predictive Coding).

11.1 Des systèmes pour faire parler les sons

11.1.1 Quelques spécimens ancestraux et actuels

Un des plus anciens systèmes permettant d'entendre un son "parler" est la *guimbarde* (cf. figure 11.1①). En Europe, on a retrouvé des exemplaires datant de l'époque gallo-romaine [8]. Plutôt que d'utiliser ses cordes vocales (la glotte[1]), on fait vibrer une lame devant sa bouche tout en articulant exagérément un mot. Avec un peu d'entraînement, on réussit à faire prononcer le mot à la vibration générée par la lame (CD, plage 21) [13].

A la fin du 18è siècle, Wolfgang von Kempelen élabora plusieurs *machines parlantes*. Des études et reconstitutions ont pu être menées [7, 16, 17] grâce à une version décrite dans [15] et un exemplaire d'époque, plus sophistiqué, conservé au *Deutsches Museum* de Munich : (a, figure 11.1②) un petit soufflet faisant office de poumons excite (c) une anche battante (glotte) débouchant sur (e,f) deux petits tuyaux (narines) puis (i) une sorte d'entonnoir en caoutchouc (bouche) duquel l'opérateur approche sa main pour former une cavité et créer différentes voyelles[2]. Avec un pilotage fin, on obtient des sons intelligibles, capables de transmettre une "intention prosodique" (CD, plage 22).

[1] La glotte est l'espace qui existe entre les bords libres des deux cordes vocales.

[2] Le système (g,h) est actionné par à-coups pour générer l'explosion de la consonne "p". Les leviers (j,k,l) sont actionnés pour reproduire respectivement les "s", "ch" et "r".

L'auteur remercie J.-S. Liénard et C. d'Alessandro pour les informations, images et sons.

Dans les années 1930, un système électro-acoustique baptisé *sonovox* consistait à mettre en appui des haut-parleurs (plus exactement, des pots vibrants) sur le cou du chanteur au voisinage de la pomme d'Adam (cf. [25]). On pouvait ainsi faire "parler" les vibrations transmises provenant, par exemple, de sons captés par un microphone[3]. Les *prothèses vocales électroniques* pour les laryngectomisés (patients ayant subi une ablation du larynx) reposent encore sur ce principe. On parle aussi d'*électrolarynx* [12].

Depuis 1970, un système baptisé *talkbox* (littéralement, "boîte parlante") est apparu[4]. Aujourd'hui plutôt utilisé par les guitaristes, il consiste à récupérer le son d'un haut-parleur avec un pavillon inversé (un entonnoir) lui-même raccordé à un tuyau. Lorsque l'extrémité libre du tuyau est placée dans la bouche, le son transmis devient, là encore, la source excitatrice du conduit vocal (cf. figure 11.1③).

Enfin, un autre système très répandu aujourd'hui sous forme électronique ou logicielle est le *vocoder*. Quels en sont les origines et le principe ?

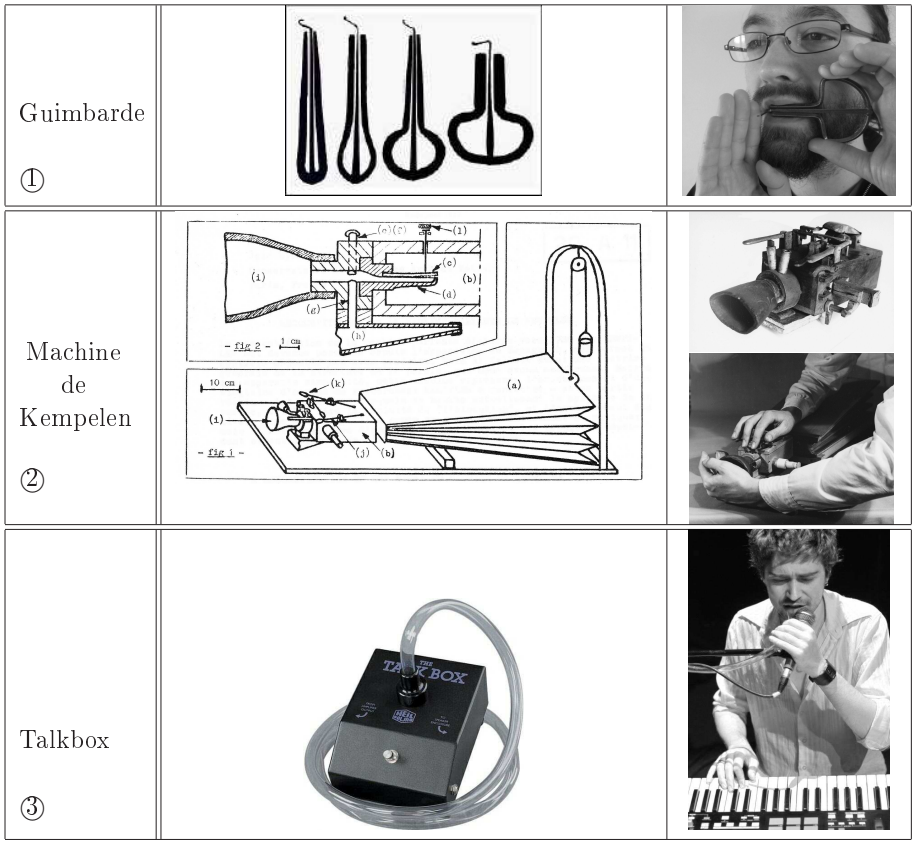

FIG. 11.1 – Illustration de différents dispositifs et de leur utilisation.

[3] Voir l'archive vidéo [3]. Mots clef (web) : *Sonovox, Kyser.*

[4] Voir des archives vidéo de premières utilisations. Mots clef : *Talkbox, Stevie Wonder.*

11.1.2 Les débuts du vocoder

Vocoder est l'abréviation de "VOice enCODER". Ce système fut initié et breveté par Homer Dudley aux Bell Labs [6] (cf. figure 11.2ⓐ). Le but ? Réduire la bande passante nécessaire à la transmission de la voix pour améliorer les débits des réseaux de télécommunications. L'idée ? Coder et reconstruire une voix en bout de chaîne à l'aide d'outils d'*analyse-synthèse* du signal. En 1928, l'*analyse* consistait à caractériser l'activité du signal de voix $v(t)$ par zones fréquentielles (comme l'affichent aujourd'hui les analyseurs de spectre de certains lecteurs audio ou chaînes hi-fi). Typiquement, un banc de filtres *passe-bande* $(F_{k=1,\dots,K})$ isolait K bandes fréquentielles desquelles on extrayait la puissance ou l'amplitude efficace correspondante $a_k(t)$.

La *synthèse* consistait à amplifier un signal d'entrée $e(t)$ dans chaque bande (F_k) par le gain associé $a_k(t)$. Pour cela, on cascadait une réplique du premier banc de filtres, K amplificateurs à gain contrôlé en tension, puis un sommateur. Avec un léger abus de langage, "on filtrait le signal $e(t)$ par une enveloppe spectrale grossière de $v(t)$ grâce à un égaliseur audio piloté".

En choisissant une entrée $e(t)$ périodique, ce système était capable de restituer des voix robotiques plutôt intelligibles qui connurent un certain succès dans le milieu du cinéma et de la musique, au contraire des télécommunications (des exemples de dispositifs commercialisés, fondés sur cette technique, sont donnés en figure 11.2 ⓑ,ⓒ,ⓓ).

FIG. 11.2 – ⓐ *Homer Dudley* aux Bell labs devant le Voder (machine parlante présentée à l'exposition universelle de New York en 1939) ; ⓑ le *Vocoder 5000* à 22 bandes d'analyse développé par EMS (Electronic Music Studios, 1976) ; ⓒ le vocoder à 16 bandes de Moog (1979) ; ⓓ le vocoder *MicroKorg* (traitement numérique à modélisation analogique) développé par Korg (2002).

En effet, pour retrouver une voix naturelle en téléphonie, il faut générer aussi un signal $e(t)$ qui reproduise les caractéristiques remarquables de l'excitation originale (émise par la glotte en général). Avec les outils de l'époque, on savait détecter si l'excitation avait une hauteur (on parle de *son voisé*), estimer et transmettre sa fréquence (appelée *pitch* ou, en chant, *fondamentale*). Alors, on générait un signal $e(t)$ de même fréquence avec une forme d'onde prédéfinie. Sinon (*son non voisé*, incluant les sons "s", "ch", "f", "t", "k", etc.), le signal $e(t)$ provenait d'un générateur de bruit. Les résultats et le coût de ce système ne furent pas assez favorables pour lancer une implantation à grande échelle[5]. En revanche, pour les applications musicales, un intérêt était justement de synthétiser des sons encore inouïs en nourrissant le système par de tels signaux $e(t)$ ou d'autres signaux étrangers à la voix.

En résumé, pour les applications musicales, le principe du vocoder est de *faire subir à un signal $e(t)$ les mêmes "renforcements fréquentiels" que ceux opérés par le conduit vocal.* Le problème se réduit donc à capturer ces *renforcements* dans le signal de voix $v(t)$ et à les reproduire sur $e(t)$.

Vocabulaire : Pour un système en général (une suspension de voiture, un circuit RLC, etc.), ces "renforcements" s'appellent les *résonances* propres du système. En parole, les résonances du conduit vocal portent le nom particulier de *formants* (cf. figure 11.3).

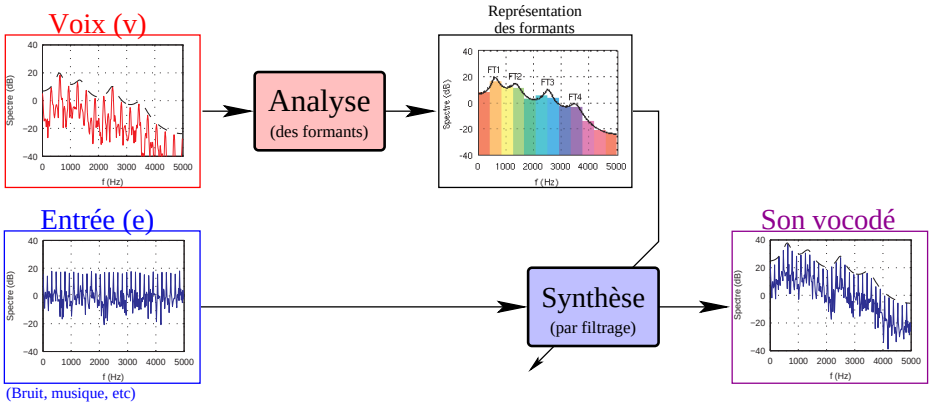

FIG. 11.3 – Principe du vocoder pour les applications musicales (vue idéalisée): ici, l'enveloppe (- -) du spectre (–) de la voix v est capturée via $K=16$ bandes (F_k, bandes de couleur) qui font apparaître 4 formants (FT1,2,3,4) ; par filtrage, cette empreinte est reproduite sur le spectre de l'entrée e pour fournir le son vocodé.

Ainsi, ces dispositifs technologiques reposent sur l'idée que *le conduit vocal est assimilable à un filtre que l'humain excite par une source* (en général, la glotte). Cette hypothèse dite de *modèle source-filtre* est-elle bien pertinente ? Vous n'êtes pas tout à fait convaincus ? Alors vérifions-là.

[5] Une première version fut intégrée dans le système de communication numérique SIGSALY de l'armée américaine. F. Roosevelt et W. Churchill l'utilisèrent pendant la seconde guerre mondiale pour communiquer outre atlantique.

11.2 Représenter la voix par un modèle source-filtre

La physique de la production de la voix est complexe. Elle met en jeu des phénomènes de mécanique des solides déformables (cartilages, muscles, tissus), de mécanique des fluides (jet, turbulences) incluant la propagation acoustique. Sa modélisation précise est encore un sujet de recherche très actif. Aussi, la représenter par un *modèle source-filtre* (cf. figure 11.4) est une hypothèse réductrice mais bien commode si elle suffit à capturer l'information recherchée (ici, reconnaître et reproduire un mot prononcé).

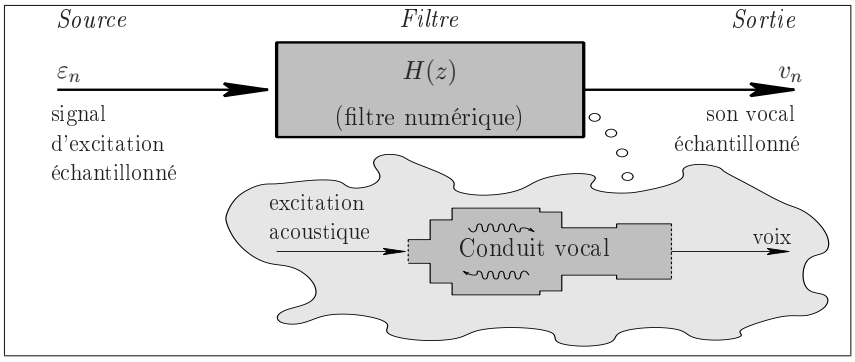

FIG. 11.4 – Dans le modèle *source-filtre*, la *source* est un générateur idéal de signal que le *filtre* transforme en "le son émis et articulé". Le filtre est destiné à prendre en charge l'action acoustique opérée par le conduit vocal. Ici, on a choisi une représentation du modèle à temps discret (le filtre est numérique et les signaux sont échantillonnés) plutôt qu'à temps continu.

Ce modèle vise à séparer les signaux d'excitation (*source* du son) du processus de transformation (*filtre*) qui variera pour créer l'articulation. Sa pertinence est, selon les cas, plutôt de type "physique", "signal" ou "perceptif".

La *source* peut représenter des signaux de nature et de provenance diverses : signaux quasi-périodiques générés par la glotte en vibration (sons voisés, chantés), bruit généré par une constriction de la glotte (voix chuchotée) ou du conduit vocal (sifflante "s", chuintante "ch"), impulsions générées par une occlusion brève du conduit (occlusive labiale "p", dentale "t", etc.), des combinaisons (sifflante "z", chuintante "j", occlusive labiale "b", dentale "d"), et bien d'autres cas. Se pose alors la question :

(**Q1**) ***Comment regrouper une telle diversité et représenter le signal source ?***

De son côté, le *filtre* doit prendre en charge et isoler les autres aspects acoustiques permettant par exemple de distinguer un "u" d'un "a", même lorsque les sources sont identiques. Une seconde question naturelle est alors :

(**Q2**) ***Peut-on isoler et exploiter des propriétés du filtre que n'a pas la source ?***

Enfin, puisque les *caractéristiques* du filtre et de la source varient pendant l'enchaînement des phonèmes, une troisième question cruciale est :

(**Q3**) ***En-dessous de quelle durée ces caractéristiques sont-elles quasi-statiques ?***

11.2.1 Quelques expériences parlantes... à faire soi-même

Enregistrez votre voix sur ordinateur. Vous avez un bon microphone ? Parfait ! Il est intégré et de piètre qualité ? Rassurez-vous, l'information voulue sera là ! D'ailleurs, prenons des conditions de téléphonie basique : réglez la fréquence d'échantillonnage[6] de votre enregistreur à $F_e = 8\,\text{kHz}$ pour réduire la plage utile à [0; 4kHz] (théorème de Nyquist-Shannon). Qu'allez-vous enregistrer ? Voici 3 expériences mais vous pourrez en décliner bien d'autres.

Expérience (E1) Enchaînez une succession de voyelles que vous chanterez sur une seule note maintenue, la plus stable possible.
(expérience a priori simple mais le besoin de concentration est quasi-garanti)

Expérience (E2) Choisissez et maintenez une voyelle que vous chanterez tout en variant à souhait la hauteur de la note. Par exemple, vous pourrez exécuter un glissando montant, descendant, un vibrato, etc.
(une fois la voyelle choisie, attention à éviter tout mouvement du conduit vocal, en particulier de la langue ou la mâchoire. Le chanteur a naturellement ce réflexe pour faciliter l'émission des notes mais ce n'est pas le but visé ici)

Expérience (E3) Reprenez l'expérience (E1) avec une voix chuchotée.

Intérêt des expériences Dans l'expérience (E1), en maintenant la note la plus stable possible, vous contrôlez la source pour la rendre la plus stationnaire possible. Les seules modifications sont dues aux voyelles prononcées. En analysant l'enregistrement, on espère donc repérer certaines caractéristiques invariantes (à relier à la source et à la question (Q1)) et des variations synchronisées avec les voyelles (à relier au filtre et à la question (Q2)). L'expérience (E2) est une version symétrique de (E1) : cette fois-ci, les variations sont reliées à la source et à (Q1), les caractéristiques du filtre sont figées. L'expérience (E3) permettra de confirmer ou rectifier les relations inférées ci-dessus. Elle permettra aussi, par comparaison avec (E1), de caractériser un autre type de source (signal bruité) pour préciser la réponse à (Q1).

11.2.2 Observations

Un enregistrement des expériences (E1-E3) est disponible sur le CD cijoint (plage 23). La note choisie pour (E1) est un ℓa_1 ($A2$ pour les anglo-saxons) de fréquence fondamentale $f^* = 110\,\text{Hz}$. L'évolution temporelle et le contenu fréquentiel des signaux sont analysés en figure 11.5 à l'aide d'un *spectrogramme*[7] pour une fenêtre de pondération de type *Hann*. Une séparation assez fine des composantes harmoniques ($f_k = k\,f^*,\ k = 1, 2, 3, \dots$) est assurée en choisissant une durée d'environ 8 périodes ($T = 70\,\text{ms} \approx 8/f^*$) pour régler la taille de la fenêtre. Que se dégage-t-il de ce spectrogramme ?

[6] contre 44.1 kHz pour le CD et 48 kHz, 96 kHz ou 192 kHz en format professionnel.

[7] Outil disponible dans Matlab (`specgram`), Scilab (`mapsound`) et d'autres logiciels parfois temps-réel. Signalons aussi le logiciel libre *PRAAT* dédié à la voix (cf. [2]).

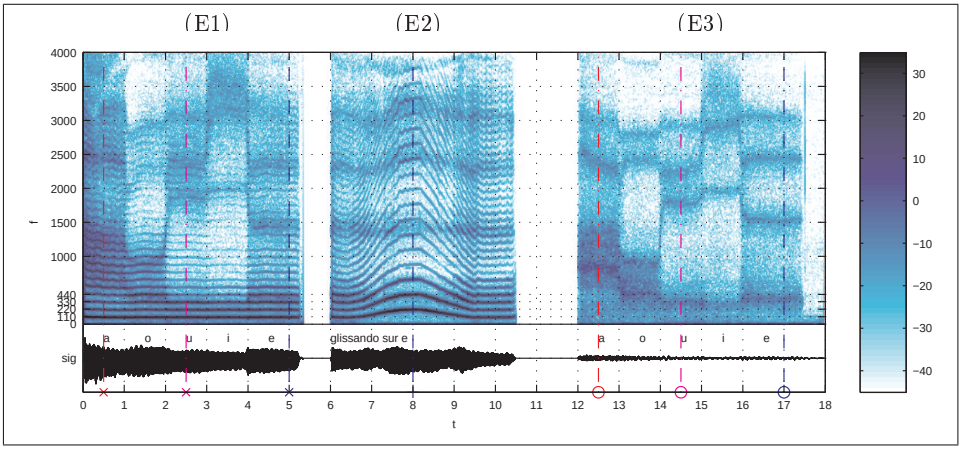

FIG. 11.5 – Spectrogramme (échelle des couleurs en décibels (dB)) et signal temporel du son (CD, plage 23) : sur [0s, 6s[, la séquence (E1) correspond à un enchaînement cadencé à 1 seconde des voyelles "a" (phonème /a/), "o" /o/, "u" /y/, "i" /i/, "e" /∅/ chantées sur la note ℓa_1 (110 Hz). Sur [6s, 12s[, la séquence (E2) correspond à la voyelle "e" /∅/ pour un glissando partant du ℓa_1 ($110\,Hz$), montant au ℓa_2 ($220\,Hz$), puis redescendant au La_1 ($110\,Hz$). Sur [$12\,s$, $18\,s$[, la séquence (E3) reprend l'enchaînement de (E1) avec une voix chuchotée plutôt que chantée.

Observation 1 Tout d'abord, dégradons la résolution de l'image (éloignez-vous de la figure) ainsi que sa luminosité (plissez les yeux). Que voyons-nous ? Pour chaque voyelle, apparaît un motif composé de (au plus) 4 bandes horizontales, superposées, épaisses et foncées. Ces 4 bandes sont les zones les plus énergétiques. Elles correspondent à 4 formants qui caractérisent les voyelles sur la plage fréquentielle $[0, 4\,kHz]$. Nous observons que les séquences de gauche (E1) et de droite (E3) dessinent le même enchaînement de motifs, ce qui correspond au même enchaînement de voyelles. La séquence (E2) fait apparaître un seul motif de 4 bandes, bien alignées avec les dernières bandes de (E1) : il s'agit de la même voyelle "e".

Observation 2 Regardons maintenant les détails du spectrogramme. La séquence (E1) est en fait constituée d'une grande quantité de rayures horizontales fines noires (dans l'ordre des fréquences croissantes : $110\,Hz$, $220\,Hz$, $330\,Hz$, etc.). Il s'agit de la décomposition en fréquences harmoniques de la note chantée. Si pour une voyelle choisie, le signal était idéalement périodique et de durée infinie, ceci correspondrait exactement à la décomposition en série de Fourier : (les rayures noires n'auraient idéalement plus d'épaisseur et deviendraient des distributions de Dirac). Sur le glissando de la séquence (E2), cette décomposition apparaît encore clairement : les rayures correspondent encore à des fréquences en relation harmonique $f_k = kf^*$ pour une fréquence instantanée f^* qui varie de $110\,Hz$ à $220\,Hz$ puis revient à $110\,Hz$. Enfin, sur (E3), aucune structure harmonique n'apparaît : le détail ressemble plutôt à un "fourmillement", autrement dit, du bruit.

Finalement, nous observons que la "partie détaillée" du spectre porte une information caractéristique de la source (structure fine harmonique, bruitée, etc.) alors que les "larges bandes fréquentielles" dans lesquelles l'énergie est concentrée (les formants) caractérisent les voyelles prononcées : l'idée de départ est retrouvée. Comment modéliser ce type d'information simplement ?

11.2.3 Hypothèses minimales et modèle paramétrique

Voici des hypothèses qui répondent (à rebours) aux questions (Q1-Q3) de manière aussi minimale que possible.

Hypothèse de stationnarité Sur la figure 11.5, les formants de (E1,E3) et les harmoniques de (E1) forment des trajectoires de fréquence quasi-constante sur des durées d'une seconde : on a un enchaînement de 6 signaux quasi-stationnaires de durée $T = 1\,s$. Pour (E2), l'hypothèse de stationnarité est encore acceptable sur des durées $T \approx 100\,ms$ car le glissando est assez lent. En général, les variations sont plus rapides (cas des occlusives "p", "d", "t", etc.). La durée T utilisée traditionnellement et qui répond à (Q3) est :

(H3) ***Le signal vocal est supposé stationnaire pour une durée de*** $\boldsymbol{T \approx 20}\,ms.$

Filtre paramétrique représentant le conduit vocal Sur une trame de durée T, les formants engendrés par le conduit vocal sont supposés immobiles. Du point de vue du signal, on les représente par les résonances d'un filtre tout-pôle d'ordre $2K$ où K correspond au nombre maximal de résonances encodables, via K paires de pôles complexes conjugués. Pour des versions échantillonnées de la source excitatrice $\{\varepsilon_n\}_{n\in\mathbb{Z}}$ et du signal vocal $\{v_n\}_{n\in\mathbb{Z}}$ (cf. figure 11.4) , le filtre numérique est représenté dans le *domaine en* $\mathcal{Z}$ par sa fonction de transfert $H(z) = \mathcal{Z}(v_n)/\mathcal{Z}(\varepsilon_n)$ donnée par

$$H(z) = G\left[1 + \sum_{k=1}^{2K} a_k\, z^{-k}\right]^{-1}, \tag{11.1}$$

où $\mathcal{Z}(x_n) = \sum_{n\in\mathbb{Z}} x_n z^{-n}$ (pour tout nombre complexe z appartenant à la couronne de convergence de la série) définit la transformée en $\mathcal{Z}$ de $\{x_n\}_{n\in\mathbb{Z}}$ et où z^{-1} représente un retard d'un échantillon. Les paramètres a_k fixent l'emplacement des pôles p_k, liés par $\prod_{k=1}^{2K}(1-p_k\,z^{-1}) = 1+\sum_{k=1}^{2K} a_k z^{-k}$. Un tel filtre (11.1) est représenté sur la figure 11.3 (enveloppe (--) du spectre de voix) pour $K=4$ paires de pôles complexes conjugués[8]. Le paramètre de gain G peut être fixé sans perte de généralité, quitte à reporter son effet sur la source excitatrice en choisissant $G\varepsilon_n$ plutôt que ε_n : on le fixe ici à $G=1$.

[8] Les paramètres sont $p_k = \overline{p_{K+k}} = \rho_k Z_{F_e}(f_k)$, avec $\rho_1 = 0,94$, $f_1 = 480\,Hz$, $\rho_2 = 0,91$, $f_2 = 1080\,Hz$, $\rho_3 = 0,93$, $f_3 = 2\,kHz$, $\rho_4 = 0,9$, $f_4 = 2,8\,kHz$. On a tracé $|H\big(Z_{F_e}(f)\big)|$ (dB) sur $[0, \frac{F_e}{2}]$ ($F_e = 8\,kHz$) où $\big(Z_{F_e}(f)\big)^{-1} = \exp(-2i\pi f/F_e)$ représente un retard de $T_e = \frac{1}{F_e}$.

En explicitant la relation $\mathcal{Z}(v_n) = H.\mathcal{Z}(\varepsilon_n)$ dans le domaine temporel, on a :

(**H2**) ***Le conduit vocal est modélisé par un filtre numérique résonant (tout-pôle) d'ordre*** $2K$ ***décrit dans le domaine temporel par l'équation de récurrence***

$$v_n + \sum_{k=1}^{2K} a_k \, v_{n-k} = \varepsilon_n. \tag{11.2}$$

Modèle paramétrique de la source Répondre à (Q1) semble relever du défi si l'on doit traiter tous les types de signaux "source" possibles. On privilégie ici une approche simple reposant sur des hypothèses communes à l'ensemble des signaux. On suppose que tout signal source est : (i) de puissance finie, (ii) de moyenne nulle[9]. Si l'on se limite à ces hypothèses, on sait que la loi de probabilité qui apportera le plus d'information (*entropie* maximale) est la loi normale $\mathcal{N}(0, \sigma^2)$. La variance σ^2 représente ici la puissance d'échantillon de la source observée en moyenne. D'après (H3), elle est supposée constante sur la trame, de sorte que :

(**H1**) ***La source échantillonnée*** ε_n ***est représentée par une variable aléatoire gaussienne*** $\boldsymbol{E}$ ***de moyenne nulle et d'écart type*** σ***, régie par la loi de probabilité***

$$p_E(\varepsilon_n|\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\big(-\frac{(\varepsilon_n)^2}{2\sigma^2}\big) \quad (\textit{avec}\ \sigma > 0), \tag{11.3}$$

où on adopte la notation standard $p(\varepsilon_n|\sigma)$ ***pour*** la densité de probabilité conditionnelle ***d'apparition de la valeur*** ε_n ***lorsqu'on est informé de la valeur*** σ***.***

Cette hypothèse sera un peu affinée en fin de chapitre.

En résumé, le modèle source-filtre retenu ici rassemble les hypothèses

$$\mathcal{H} = \{H1, H2, H3\}. \tag{11.4}$$

Ce modèle permet d'encoder jusqu'à K formants ainsi que la puissance du signal de voix pour chaque trame de signal stationnaire de durée $T = N/F_e \approx 20\,ms$ (N est le nombre d'échantillons). Les paramètres T et K seront choisis pour régler le vocoder. L'analyse du son vocal consistera à estimer pour chaque trame les coefficients du filtre et la puissance de la source, c'est-à-dire, le vecteur

$$\Theta = (\, a_1, \ldots, a_{2K}\,, \sigma\,)^t. \tag{11.5}$$

Maintenant que le modèle *source-filtre* est décrit par $\mathcal{H}$, se pose la question :

(**Q4**) ***Comment estimer*** Θ ***à partir d'une trame de signal vocal échantillonné ?***

[9] On s'intéresse à des signaux acoustiques centrés autour d'un état d'équilibre (pression atmosphérique, débit moyen de l'écoulement d'air, etc).

L'estimation proposée ici est celle dite "par LPC". Avant de la présenter, voici un dernier paragraphe qui éclaire (H2) par quelques notions d'acoustique.

Considérations acoustiques et ordres de grandeurs Représentons de façon approchée un conduit vocal de longueur L par une cascade de K tubes cylindriques de longueurs identiques L/K, à l'intérieur desquels se propagent des ondes planes (cf. figure 11.4 en bas). De même, représentons la glotte par un générateur idéal de débit et l'effet du rayonnement aux lèvres par une impédance acoustique de faible valeur. Que pouvons-nous tirer d'une telle description caricaturale ? La durée du voyage d'une onde d'une extrémité à l'autre d'un tronçon vaut $\tau = L/(Kc)$ ($c \approx 340\,m/s$ est la célérité du son). À chaque jonction, la continuité de la pression et du débit implique qu'une fraction de l'onde (dépendant du rapport des sections de tube) est réfléchie dans le tronçon tandis que la partie complémentaire est transmise (le phénomène aux extrémités est similaire). Les combinaisons sur l'ensemble des jonctions conduisent à une fonction de transfert entre signaux acoustiques (e.g. débit glottique vers débit aux lèvres) donné par $H(z)\,z^{-K}$ où H est de la forme (11.1) et où z^{-1} correspond à un retard de durée τ. On retrouve donc (H2) dans la mesure où le retard[10] z^{-K} n'a pas d'effet sur les résonances.

Donnons maintenant quelques ordres de grandeurs. La longueur typique d'un conduit vocal adulte masculin est $L = 17\,cm$. En réglant τ comme la période d'échantillonnage téléphonique standard ($\tau = 1/F_e$, $F_e = 8\,kHz$), on trouve que le nombre de tronçons vaut $K = L/(c\tau_e) = LF_e/c = 4$. Du point de vue du signal (cf. (H2)), K est aussi le nombre maximal de formants encodables par H. Ceci explique donc le nombre de formants que nous avons observés sur le spectrogramme (figure 11.5) dans la plage fréquentielle $[0, F_e/2[= [0, 4\,kHz[$.

Terminons par deux points sur la validité du modèle acoustique. Premièrement, la cascade de tubes droits néglige la dérivation vers le nez. Lorsqu'elle communique (voyelles nasalisées "en", "on", "in", etc), la cavité nasale prend de l'énergie au conduit et crée essentiellement des anti-résonances. Ces zéros pourraient être ajoutés via un numérateur $N(z)$ dans (11.1) : vous testerez si le vocoder arrive à capturer la nasalisation sans ajouter cette complication. Deuxièmement, la description en ondes planes est acceptable pour des longueurs d'ondes $\lambda = c/f$ suffisamment grandes devant celles des premiers modes transverses. Donnons un ordre de grandeur dans le cas simple d'une section rigide rectangulaire de grand côté ℓ : le mode transverse le plus bas (vitesse transverse en demi-arche de sinusoïde, nulle sur la paroi) correspond à $\lambda = 2\ell$. Le cas le plus défavorable ($\ell \approx 4\,cm$) conduit à $f_{critique} = 340/(2\times 4e-2) = 4250\,Hz$. Ainsi, le choix $F_e/2 = 4\,kHz$ choisi en téléphonie correspond à la valeur limite qui permet de capturer les formants explicables par un modèle de propagation acoustique longitudinale.

[10] Durée $K\tau = L/c$ de propagation d'une distance L (ici, glotte-lèvres), soit K tronçons.

11.3 Méthode par LPC (Linear Predictive Coding)

Le sigle LPC ne vous parle pas ? Sa signification de "codage par prédiction linéaire" ne vous suggère aucun lien évident avec ce qui précède ? Cette perplexité s'explique : le moment est venu d'élucider le titre du chapitre.

Sigle LPC Plutôt que de s'appuyer sur des notions de l'acoustique de la voix, de l'analyse temps-fréquence de son signal v_n, ou du modèle source-filtre, le sigle LPC se réfère à des notions de *prédiction* et de *codage*. Comme pour le mot *vocoder*, ce vocabulaire est hérité de la téléphonie et de ses préoccupations. Lesquelles ? Trouver un prédicteur $\widehat{v}_n$ de l'échantillon v_n. Comment ? À l'aide une fonction simple des échantillons passés $v_{n'\leq n-1}$: une combinaison linéaire. Le rideau se lève et entre en scène le *prédicteur linéaire*

$$\widehat{v}_n = \sum_{j=1}^{J} \alpha_j v_{n-j}, \tag{11.6}$$

où $\alpha_{j=1,\dots,J}$ sont les J coefficients de prédiction. Pourquoi ? Comme pour le vocoder de H. Dudley, le but était de réduire le débit nécessaire à la transmission (cette fois-ci, numérique) de la voix. L'idée était de reconstruire de courtes portions de M échantillons à l'aide des J coefficients α_j transmis, plus quelques autres ([24],[9, §8]), dans le cas $J \ll M$. Ce *codage* du signal par les coefficients d'un *prédicteur linéaire* est ainsi à l'origine du sigle LPC. Aujourd'hui, il est une brique de base de la plupart des codeurs de voix[11].

Lien avec le modèle source-filtre $\mathcal{H}$ Dans l'approche par LPC, on choisit les α_j pour que sur la portion des M échantillons traités, l'*erreur de prédiction*

$$\eta_n = v_n - \widehat{v}_n \tag{11.7}$$

soit minimale en un sens à préciser[12]. Pour les coefficients optimaux, cette erreur porte le nom d'*innovation* : quand le processus n'est pas exactement prédictible ($\eta_n \neq 0$), on le considère comme doué d'innover spontanément. Un lien évident entre (11.6)-(11.7) et (11.1) apparaît en choisissant d'identifier :

$$J=2K, \quad \eta_n=\varepsilon_n, \quad \alpha_j=-a_j, \quad \text{et} \quad M=N \quad \text{(pour rappel, } T=N/F_e\text{)}.$$

L'équivalence semble miraculeuse pour des approches aussi différentes (considérations *a priori* sur un signal abstrait pour le prédicteur contre l'exploitation de spécificités de la voix pour $\mathcal{H}$). Cette rencontre ne tiendrait probablement plus si l'on devait raffiner ces approches. Il est néanmoins remarquable que les vocables *innovation* et *excitation* trouvent des interprétations concordantes : sur une trame, l'*excitation* du conduit vocal est bien la partie qui varie et *innove* le plus dans le processus de génération du signal de voix.

[11] La toute première utilisation eut lieu en décembre 1974 entre Culler-Harrison Incorporated (Gioleta, Californie) et le laboratoire Lincoln du MIT (Lexington, Massachuetts) [9].

[12] Par exemple, l'erreur quadratique moyenne apparaît comme un choix intuitif et simple.

Le lien avec (H2,H3) explique le vocabulaire. Reste à construire un estimateur de Θ fondé sur $\mathcal{H}$ (en particulier (H1)) plutôt que sur la minimisation d'une erreur moyenne abstraite d'un prédicteur donné *a priori*. Entre en scène, la *vraisemblance* des données observées. De quoi s'agit-il ?

11.3.1 Approche retenue pour l'analyse du signal de voix

Isolons une trame de signal $(v_1, v_2, \ldots, v_N)$ observé pendant une durée $T = N/F_e$. Supposons ces échantillons correctement décrits par $\mathcal{H}$. Estimer la valeur de Θ pour ces données, c'est se poser la question :

(Q4) ***Quelle valeur de*** Θ ***concorde le mieux avec la trame*** $(v_1, v_2, \ldots, v_N)$ ***?***

Quelle signification attribuer à "concorder le mieux" ? Un choix sensé est de considérer qu'il s'agit de la valeur de Θ pour laquelle *les données observées auront la plus grande probabilité d'apparaître*. En d'autres termes, il s'agit de :

Étape (i) : modéliser $(v_1, v_2, \ldots, v_N)$ comme *une réalisation* de N variables aléatoires sous la connaissance du modèle de génération $\mathcal{H}$ et Θ ;

Étape (ii) : rendre maximale leur probabilité d'apparition selon Θ.

La densité de probabilité de $(v_1, \ldots, v_N)$ connaissant $\mathcal{H}$ et Θ est notée

$$P_N = p\big(\, v_1, \ldots, v_N \,\big|\, \mathcal{H}\,,\, \Theta\,\big). \tag{11.8}$$

Elle définit la **fonction de vraisemblance** des données $v_1, \ldots, v_N$ pour $\mathcal{H}$ et Θ.

La découverte de cette fonction intrigue parfois lorsqu'on s'aperçoit qu'il s'agit d'une probabilité conditionnelle pour laquelle : (a) les *variables aléatoires* correspondent à la *partie connue* (ici, les données $v_1, \ldots, v_N$) et (b) la *partie inconnue* est considérée *déterministe* (ici, Θ). En fait, ceci n'a rien de contradictoire. La nature *aléatoire* ou *déterministe* des variables provient du *choix d'un modèle*[13] (ici, $\mathcal{H}$). Par ailleurs, le statut *connu* ou *inconnu* des mêmes variables ne dépend pas du choix de modèle mais du *choix d'une expérience*[14] (ici, la captation d'un signal vocal).

Le principe du **maximum de vraisemblance** a justement l'originalité d'établir une valeur pertinente d'une variable déterministe inconnue à partir de l'observation d'une réalisation de variables aléatoires. Il est retenu ici :

(H4) ***La valeur optimale du paramètre*** Θ ***pour une trame*** $v_1, \ldots, v_N$ ***donnée est celle qui rend maximale la vraisemblance (11.8) de cette trame.***

[13] Le lancé de dé est un cas emblématique : la mauvaise maîtrise et la sensibilité aux conditions expérimentales (géométrie, matériaux, conditions initiales) amènent à représenter le numéro obtenu par une variable aléatoire plutôt que l'équilibre de fin de trajectoire gouvernée par des lois de mécanique. Cette préférence se justifie mais reste un *choix de modèle*.

[14] Vous voici phylogénéticien ! Après analyse au microscope, un minuscule échantillon pileux collecté par vos soins semble appartenir à une espèce non identifiée. Il pourrait s'agir d'un fameux chaînon manquant dans la théorie de l'évolution de Darwin. Pas question de recourir à de l'aléatoire ! Pour votre prix Nobel, il vous faut deux preuves : l'âge (datation au carbone 14) et le génome (réplication d'ADN à partir d'une solution protéinée récente). Ces technologies requièrent et détruiront chacune l'échantillon complet ! C'est l'expérience choisie (impérativement unique) qui sélectionnera la quantité connue et celle inconnue...

11.3.2 Résolution

Étape (i) L'expression de la vraisemblance se déduit des 2 règles de calcul de probabilités, rappelées ci-dessous :

(PC) Probabilité conditionnelle : la probabilité d'apparition des évènements a et b sachant c est donnée par

$$p(a,b|c) = p(a|b,c)\,p(b|c) = p(b|a,c)\,p(a|c).$$

(CV) Changement de variable : les densités de probabilité p_Y et p_X de deux variables aléatoires en bijection $y=f(x)$ sont reliées par

$$p_Y(y) = \big|f'\circ f^{-1}(y)\big|^{-1}\ p_X\big(f^{-1}(y)\big).$$

Cette formule se généralise pour des probabilités conditionnées par une hypothèse commune $\widetilde{H}$, simplement avec $p_X(x|\widetilde{H})$ et $p_Y(y|\widetilde{H})$.

D'après (PC), en partitionnant $(v_1,\dots,v_n)$ en $a=(v_1,\dots,v_{n-1})$ et $b=v_n$, on trouve la relation générale suivante pour $2\le n\le N$,

$$P_n = p(v_1,\dots,v_n\,|\,\Theta,\mathcal{H}) = \underbrace{p(v_n\,|v_1,\dots,v_{n-1},\Theta,\mathcal{H})}_{\overset{\text{déf}}{=}\pi_n}\ \underbrace{p\big(v_1,\dots,v_{n-1}\,|\,\Theta,\mathcal{H}\big)}_{=P_{n-1}}, \tag{11.9}$$

et $P_N=\pi_N\,P_{N-1}=\pi_N\,\pi_{N-1}\,P_{N-2}=\cdots=\pi_N\dots\pi_2\,P_1$ où $P_1=\pi_1$. Reste à exprimer π_n (vraisemblance de v_n pour $v_1,\dots,v_{n-1},\Theta,\mathcal{H}$) grâce à notre modèle.

Dans (CV), on a $p_Y=\pi_n$ en choisissant $y=v_n$ et $\widetilde{H}=\{v_1,\dots,v_{n-1},\Theta,\mathcal{H}\}$. La densité de probabilité donnée par (H1) est celle de $x=\varepsilon_n$. Le changement de variable fourni par (H2) s'écrit $y=f(x)=x+\kappa$ avec $\kappa=-\sum_{k=1}^{2K}a_k v_{n-k}$. Pour que f définisse bien une fonction et pour appliquer (CV), κ doit être connue sous l'hypothèse $\widetilde{H}$. Cette condition est remplie si $n>2K$. Ainsi, on a

$$\begin{aligned}\pi_n &= p(v_n\,|\,v_1,\dots,v_{n-1},\Theta,\mathcal{H}) = p_E\left(v_n+\sum_{k=1}^{2K}a_k v_{n-k}\,\middle|\,v_1,\dots,v_{n-1},\Theta,\mathcal{H}\right)\\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\big[v_n+\sum_{k=1}^{2K}a_k v_{n-k}\big]^2}{2\sigma^2}\right), \qquad \text{si } n>2K. \end{aligned}\tag{11.10}$$

Au final, appliquer (11.9-11.10) pour $n\in\mathbb{T}=[2K{+}1,N]_{\mathbb{N}}$ conduit à une forme semi-explicite $P_N=f_{\mathbb{T}}(\Theta)\,P_{2K}$ où $f_{\mathbb{T}}(\Theta)=\prod_{n\in\mathbb{T}}\pi_n$ est donnée par

$$f_{\mathbb{T}}(\Theta) = \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^{d(\mathbb{T})}\exp\Big(-\frac{Q_{\mathbb{T}}(\Theta)}{2\sigma^2}\Big)\ \text{ avec }\ Q_{\mathbb{T}}(\Theta)=\sum_{n\in\mathbb{T}}\Big[v_n+\sum_{k=1}^{2K}a_k v_{n-k}\Big]^2, \tag{11.11}$$

avec $d(\mathbb{T})=\operatorname{card}\mathbb{T}=N-2K$ et $\Theta=(a_1,\dots,a_{2K},\sigma)^t$. Poursuivre le procédé itératif pour $n\le 2K$ requiert une information absente de $\mathcal{H}$ (de v_0 pour π_{2K} à v_{-2K+1} pour π_1) récusée par (H3). Expliciter P_{2K} pose donc la question :

(Q5) ***Quelle information considérer en bord de trame ?***

Escale : la méthode par LPC, sa jungle, ses choix Le problème qui émerge ici n'est pas anodin. Il révèle que (H4) n'apporte pas de réponse bien posée à (Q4) pour le modèle $\mathcal{H}$. L'ouvrage [9, partie 1] de Robert Gray[15] sur l'histoire de la méthode en témoigne : (Q4) n'a pas de réponse univoque évidente. D'ailleurs, certaines approches ne sont ni fondées sur la vraisemblance, ni sur $\mathcal{H}$ et certaines variantes règlent même d'emblée le problème de gestion du bord de trame. Il n'en reste pas moins que, dans tous les cas, de manière explicite ou implicite, un choix est fait qui répond à (Q5). Des solutions standard reviennent, du point de vue de la vraisemblance, à :

(A) s'intéresser encore à la vraisemblance de la trame $(v_1, \dots, v_N)$, quitte à compléter $\mathcal{H}$ par une information annexe (*a priori* ou *contextuelle*),

(B) s'intéresser plutôt à la vraisemblance d'un *signal virtuel idéalisé* $(v_n)_{n\in\mathbb{Z}}$, dont la trame ne représenterait qu'un extrait.

Voici deux exemples (1 et 2) pour chacune des deux approches (A et B).

(H5) ***On modifie (H4) soit en complétant $\mathcal{H}$ par (A1) ou (A2), soit en considérant la vraisemblance des signaux $(v_n)_{n\in\mathbb{Z}}$ décrits par (B1) ou (B2) avec***

choix (A1) : ***les échantillons $v_{1-2K}, \dots, v_0$ sont considérés nuls,***

choix (A2) : ***les échantillons $v_{1-2K}, \dots, v_0$ sont connus (car mesurés),***

choix (B1) : ***la trame $(v_1, \dots, v_N)$ isole l'extrait non nul du signal complet,***

choix (B2) : ***la trame est extraite d'un signal stationnaire ergodique[a].***

[a] Rappel : un processus aléatoire est stationnaire si ses propriétés statistiques sont indépendantes du temps. Il est de plus ergodique si ses moyennes statistiques (construites avec l'espérance mathématique) sont égales aux moyennes temporelles (i.e. si les statistiques se vérifient en moyenne dans le temps).

Les approches (A) permettent le calcul itératif de P_{2K} qui conduit à (11.11) avec $\mathbb{T} = [1, N]_{\mathbb{N}}$ où, pour tout $n < 1$, v_n est nul (cas A1) ou donné (cas A2).

Le cas (B1) interprète $(v_n)_{n\in\mathbb{Z}}$ comme un silence interrompu par une partie courte de phonème. Une propriété immédiate de (B1) est que pour tout $\mathbb{T} \supseteq \mathbb{T}^\star = [1-2K, N+2K]_{\mathbb{N}}$, la vraisemblance de $(v_n)_{n\in\mathbb{T}}$ vaut $f_{\mathbb{T}}(\Theta)$ avec

$$Q_{\mathbb{T}}(\Theta) = \sum_{n\in\mathbb{Z}} \Big[v_n + \sum_{k=1}^{2K} a_k v_{n-k}\Big]^2 = \sum_{n\in\mathbb{T}^\star} \Big[v_n + \sum_{k=1}^{2K} a_k v_{n-k}\Big]^2 \quad \Big(= \sum_{n\in\mathbb{T}^\star} \varepsilon_n^2\Big).$$

Ainsi, les coefficients optimaux a_k et le filtre H associé restent invariants pour tout $\mathbb{T} \supseteq \mathbb{T}^\star$. Ce n'est pas le cas de la puissance moyenne σ^2 : ici, le paramètre invariant (pour $\mathbb{T} \supseteq \mathbb{T}^\star$) est l'énergie totale de la source $\sum_{n\in\mathbb{T}^\star} \varepsilon_n^2$.

La particularité de (B2) est de chercher à n'apporter *aucune information extérieure* à la trame $(v_1, \dots, v_N)$: on préfère aboutir à la forme $f_{\mathbb{T}}(\Theta)$ en préservant des propriétés statistiques plutôt qu'en imposant des valeurs de signal, pour des raisons et en un sens reprécisés plus loin.

Ces choix conduisent tous à la forme $f_{\mathbb{T}}$ et ont peu d'impact numérique dès que $N \gg 2K$ (typiquement, $N = 8000 \times 0.02 = 160, 2K = 8$). Leur motivation vient plutôt de propriétés. Pour fixer un choix, intéressons-nous à l'étape (ii).

[15] L'un des pères de la méthode par LPC.

Étape (ii) *Maximiser la vraisemblance* $f_{\mathbb{T}}$ selon Θ (obtenue avec (11.11) pour un choix $(v,\mathbb{T})$ de (H5)), c'est encore *minimiser l'anti-log-vraisemblance*

$$L_{\mathbb{T}} = -\ln f_{\mathbb{T}} = \frac{d(\mathbb{T})}{2}\ln(2\pi) + d(\mathbb{T})\,\ln\sigma + \frac{Q_{\mathbb{T}}}{2\sigma^2}, \tag{11.12}$$

puisque $x \mapsto -\ln x$ est strictement décroissante (éliminer ainsi l'exponentielle de $Q_{\mathbb{T}}$ simplifie les calculs de l'optimisation). Ceci ramène à la question :

(**Q6**) ***Quelle valeur de*** Θ ***minimise l'anti-log-vraisemblance*** $L_{\mathbb{T}}$ ***?***

Comme $\Theta \mapsto L_{\mathbb{T}}(\Theta)$ est régulière[16], son gradient s'annule à ses extrema :

$$\frac{\partial L_{\mathbb{T}}}{\partial \Theta} = (0,\ldots,0)^t \in \mathbb{R}^{2K+1}, (\text{ avec pour rappel, } \Theta = (a_1,\ \ldots\ , a_{2K}\ ,\ \sigma\ )^t).$$

On examine la dérivée partielle selon σ d'une part, et les a_k d'autre part. Ainsi,

$$\frac{\partial L_{\mathbb{T}}}{\partial \sigma} = \frac{d(\mathbb{T})}{\sigma} - \frac{Q_{\mathbb{T}}}{\sigma^3} = \frac{d(\mathbb{T})}{\sigma^3}\left(\sigma^2 - \frac{Q_{\mathbb{T}}}{d(\mathbb{T})}\right)$$

définit une fonction de $\sigma(>0)$ négative puis positive qui s'annule si[17]

$$\sigma^2 = \frac{Q_{\mathbb{T}}}{d(\mathbb{T})} \quad \textit{(variance optimale)}, \text{ soit encore } \sigma = \sqrt{\frac{Q_{\mathbb{T}}}{d(\mathbb{T})}}\ \textit{(écart-type)}. \tag{11.13}$$

D'autre part, l'optimalité[18] de a_p $(1 \le p \le 2K)$ est décrite par (cf. (11.11-11.12))

$$\begin{aligned} 0 &= \frac{\partial L_{\mathbb{T}}}{\partial a_p} = \frac{1}{2\sigma^2}\frac{\partial Q_{\mathbb{T}}}{\partial a_p} = \frac{1}{\sigma^2}\sum_{n\in\mathbb{T}}\left[v_{n-p}\left(v_n + \sum_{k=1}^{2K} a_k v_{n-k}\right)\right] \\ &= \frac{1}{\sigma^2}\left[R_{p,0} + \sum_{k=1}^{2K} R_{p,k}a_k\right], \text{ avec } R_{p\,,\,k} = \sum_{n\in\mathbb{T}} v_{n-p}v_{n-k}. \end{aligned} \tag{11.14}$$

Finalement, le vocoder peut être construit à partir d'un choix de (H5) et de

(**H6**) ***La valeur de*** Θ ***qui minimise*** $L_{\mathbb{T}}$ ***s'obtient avec l'algorithme suivant :***

I. Résoudre le système linéaire formé par les $2K$ ***équations (11.14),***

$$\begin{pmatrix} R_{1,1} & R_{1,2} & \ldots & R_{1,2K} \\ R_{2,1} & R_{2,2} & & R_{2,2K} \\ \vdots & & & \vdots \\ R_{2K,1} & R_{2K,2} & \ldots & R_{2K,2K} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{2K} \end{pmatrix} = - \begin{pmatrix} R_{1,0} \\ R_{2,0} \\ \vdots \\ R_{2K,0} \end{pmatrix}.$$

II. Calculer l'écart-type (11.13) grâce à (11.11) avec les a_p ***optimaux.***

En résumé, il s'agit d'inverser la matrice symétrique positive $(2K \times 2K)$ qui apparaît dans (H6.I), puis de calculer une somme de carrés dans (H6.II).

[16] On exclut les trames de signal nul, seuls cas pour lesquels σ et $Q_{\mathbb{T}}$ sont nuls.

[17] L'optimum est donc bien un minimum. Il correspond à choisir σ^2 comme l'*énergie moyenne de la source* sur $\mathbb{T}$ puisque $Q_{\mathbb{T}} = \sum_{n\in\mathbb{T}} \varepsilon_n^2$ (énergie de la source cumulée sur $\mathbb{T}$).

[18] D'après (11.11), on récrit $Q_{\mathbb{T}}(\Theta) = (1, a_1, \ldots, a_{2K})\, M\, (1, a_1, \ldots, a_{2K})^t$ où la matrice $M = M^t = \sum_{n\in\mathbb{T}} {u_n}^t u_n$ avec $u_n = (v_n, v_{n-1}, \ldots, v_{n-2K})^t$ est symétrique positive. Cette dernière propriété garantit qu'une valeur extrémale de $Q_{\mathbb{T}}$ selon les a_k sera un minimum.

Toëplitz or not Toëplitz ? Voici le problème résolu pour les choix (A1), (A2) et (B1). Vite, vous implantez votre vocoder en profitant du "canevas de code" gracieusement offert en section 11.4. Après quelques instants de labeur, vos premiers sons vocodés arrivent et tout cela vous paraît épatant. Pourtant, rajouter un dernier effort eût été bien profitable : jugez par vous-même...

L'algorithme qu'il reste à établir (algorithme de *Durbin-Levinson*) est responsable du succès technologique de la méthode par LPC qu'on ne trouve d'ailleurs que sous cette forme. Son efficacité lui a valu d'être intégré dans les premiers téléphones mobiles et autres systèmes à ressources limitées. Quelle est cette prodigieuse propriété ? En exploitant astucieusement la structure de Toëplitz[19] symétrique d'une matrice donnée, il est possible de ramener le calcul de son inverse d'une complexité algorithmique de $\mathcal{O}(P^3)$ à $\mathcal{O}(P^2)$.

Ceci nous concernerait-il ? Presque pour les cas (A) et oui pour les cas (B). En effet, considérons la matrice symétrique définie dans (H6.I) :

cas (A) : Les quantités $R_{p,k}$ et $R_{p+1,k+1}$ partagent formellement N–1 termes sur les N qui les composent (avec $\mathbb{T}=[1,N]_{\mathbb{N}}$). Pour $p,k\geq 1$, la déviation vaut $R_{p,k}-R_{p+1,k+1}=v_{N-p}v_{N-k}-v_{1-p}v_{1-k}$ (où $v_{1-p}v_{1-k}=0$ pour (A1)).

cas (B1) Ce cas conduit bien à une matrice de Toëplitz puisque (cf. p.18), pour tout $(p,k)\in[1,2K]^2_{\mathbb{N}}$ et $\mathbb{T}\supseteq\mathbb{T}^\star$, $R_{p,k}=\sum_{n\in\mathbb{T}}v_{n-p}\,v_{n-k}=\sum_{n\in\mathbb{Z}}v_{n-p}\,v_{n-k}$ est une fonction de $|p-k|$. On note $R_{p,k}=\mathcal{R}\big(|p-k|\big)$.

cas (B2) : Pour un signal aléatoire $(\widetilde{v}_n)_{n\in\mathbb{Z}}$, l'hypothèse de stationnarité implique que pour tout n et p, $\mathcal{E}(\widetilde{v}_m\widetilde{v}_n)$ ne dépend que de $|m-n|$ ($\mathcal{E}$ désigne l'*espérance*) : $\mathcal{E}(\widetilde{v}_n\widetilde{v}_{n-p})=\widetilde{\mathcal{R}}\big(|p|\big)$ définit l'auto-corrélation du signal $(\widetilde{v}_n)_{n\in\mathbb{Z}}$. L'hypothèse d'ergodicité signifie que l'espérance correspond aussi à la moyenne temporelle. Supposer (B2) revient donc à supposer que $\widetilde{v}_n$ est le signal v_n ($n\in\mathbb{T}=[1,N]_{\mathbb{N}}$) prolongé hors de $n\in[1,n]_{\mathbb{N}}$ de sorte qu'il conserve les mêmes propriétés statistiques (ici, une moyenne nulle et $R_{i,j}$ proportionnel à $\widetilde{\mathcal{R}}_{|i-j|}$ de sorte qu'on retrouve la propriété de Toëplitz). Deux estimateurs de l'auto-corrélation $\widetilde{\mathcal{R}}\big(|p|\big)$ pour une trame finie $(v_1,\ldots,v_N)$ sont donnés par (cf. e.g. [20]) :
(B2a) : $\widetilde{\mathcal{R}}_a(k)=\frac{1}{N}\sum_{n=1+|k|}^{N}v_nv_{n-|k|}$ (estimateur biaisé),
(B2b) : $\widetilde{\mathcal{R}}_b(k)=\frac{1}{N-|k|}\sum_{n=1+|k|}^{N}v_nv_{n-|k|}$ (estimateur non biaisé).

Choix final, propriétés et équations de Yule-Walker Du point de vue du traitement de signal, les choix de type (A) reviennent à appliquer une fenêtre de pondération sur le signal source seul[20]. Ils conduisent aux méthodes par LPC souvent dites "à covariance". Dans ce cas, non seulement

[19] Rappel : une matrice T est de Toëplitz si chaque diagonale est remplie par une même valeur, c'est-à-dire, si $T_{p+1,k+1}=T_{p,k}$.

[20] Ici, la source est considérée active (variance pondérée par $w_n=1$) pour $1\leq n\leq N$ et inactive (variance pondérée par $w_n=0$) sinon. Ainsi, la probabilité de sa nullité hors de la trame est certaine et on ne peut que se limiter à étudier la vraisemblance de la trame.

la matrice de (H6.I) n'est pas de Toëplitz mais il n'y a pas de garantie d'obtenir un filtre H stable. ces méthodes sont donc écartées ici[21].

Les choix de type (B) reviennent à appliquer une fenêtre de pondération sur le signal vocal seul (pour plus de détails, consulter [21], par exemple). Ils conduisent aux méthodes par LPC dites "à auto-corrélation". La matrice de (H6.I) est alors de Toëplitz. De plus, pour le choix de l'auto-corrélation biaisée, on sait que le filtre obtenu est stable à phase minimale (il transfert l'énergie au plus vite). C'est ce choix, commun à B1 et B2 à une constante multiplicative près, que nous retiendrons finalement :

(H7) ***Une trame de support*** $\mathbb{T} = [1, N]_{\mathbb{N}}$ ***est supposée extraite d'un processus stationnaire ergodique dont l'auto-corrélation est donnée par son estimateur biaisé*** $\widetilde{\mathcal{R}}_a$***. Sous cette hypothèse,*** $d(\mathbb{T}) = N$ ***et, dans (11.14) et (H6.I),*** $R_{p,k}$ ***est remplacé par l'estimateur***

$$R_{p,k} \;\approx\; N\,\widetilde{\mathcal{R}}_a\big(|p-k|\big) \;=\; \sum_{n=1+|p-k|}^{N} v_n\, v_{n-|p-k|} \;=\; \sum_{n\in\mathbb{Z}} v_n\, v_{n-|p-k|}, \qquad (11.15)$$

où l'on a prolongé le signal $v_1, \ldots, v_N$ ***par des zéros.***

La résolution de (H6) avec (H7) fournit le paramètre optimal Θ recherché. Toutefois, on préfère souvent regrouper (H6.I) et (H6.II) sous la forme d'une unique équation matricielle en remarquant que, d'après (11.11), (11.13), (11.14) et en notant $a_0 = 1$, on a

$$\begin{aligned} N\sigma^2 = Q_{\mathbb{T}} \;&=\; \sum_{k_1=0}^{2K} a_{k_1} \sum_{k_2=0}^{2K} a_{k_2} \sum_{n\in\mathbb{Z}} v_{n-k_1} v_{n-k_2} = \sum_{k_1=0}^{2K} a_{k_1} \sum_{k_2=0}^{2K} a_{k_2} R_{|k_1-k_2|} \\ &=\; \sum_{k_1=0}^{2K} a_{k_1} R_{k_1}. \end{aligned}$$

Ceci fournit la première ligne des "équations de Yule-Walker" données par :

(H8) ***On a à résoudre en*** $\Theta = (\,a_1, \ldots, a_{2K}\,, \sigma\,)^t$***,***

$$\left(\begin{array}{c|ccccccc} R_0 & R_1 & R_2 & \ldots & \ldots & \ldots & R_{2K} \\ \hline R_1 & R_0 & R_1 & \ldots & \ldots & \ldots & R_{2K-1} \\ R_2 & R_1 & R_0 & R_1 & \ldots & \ldots & R_{2K-2} \\ \vdots & \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \ddots & R_1 \\ R_{2K} & R_{2K-1} & \ldots & \ldots & \ldots & R_1 & R_0 \end{array}\right) \left(\begin{array}{c} 1 \\ \hline a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_{2K} \end{array}\right) = \left(\begin{array}{c} N\sigma^2 \\ \hline 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{array}\right) \qquad (11.16)$$

[21] En revanche, indiquons que le signal est mieux expliqué (résidu ε plus faible) par ces approches que celles de type (B).

11.3.3 Algorithme de Durbin-Levinson

Pour ramener de $\mathcal{O}(P^3)$ à $\mathcal{O}(P^2)$ la complexité, la résolution de l'algorithme de Durbin-Levinson s'appuie sur deux points clefs :

(A) l'exploitation d'une identité remarquable satisfaite par les matrices de Toëplitz symétriques (propriété 1 ci-dessous),

(B) une procédure itérative sur l'ordre P du modèle.

Propriétés 1 (Matrice de Toëplitz symétrique) *Soit J_p la matrice anti-diagonale $p \times p$ définie par*

$$J_p = \begin{bmatrix} 0 & \dots & 0 & 1 \\ \vdots & \iddots & \iddots & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix}.$$

Soit T_p une matrice de Toëplitz symétrique $p \times p$. Alors T_p vérifie l'identité

$$J_p \, T_p \, J_p \, = \, T_p. \tag{11.17}$$

Ceci se déduit immédiatement de la remarque 1(i) suivante.

Remarque 1 *(i) Multiplier une matrice par J_p à gauche (respectivement, à droite) inverse l'ordre de ses lignes (respectivement, de ses colonnes) ; (ii) La matrice $(J_p)^2$ est la matrice identité. En résumé,*

$$J_p \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_p \end{bmatrix} = \begin{bmatrix} L_p \\ \vdots \\ L_2 \\ L_1 \end{bmatrix}, \quad \begin{bmatrix} C_1 \, C_2 \, \cdots \, C_p \end{bmatrix} J_p = \begin{bmatrix} C_p \, \cdots \, C_2 \, C_1 \end{bmatrix} \quad \textit{et} \quad (J_p)^2 = I_p.$$

Problème à résoudre Considérons la suite de problèmes $(\mathcal{P}_p)_{p=0,\dots,P}$ avec

$$(\mathcal{P}_p): \quad \mathcal{R}_p \begin{bmatrix} 1 \\ \mathcal{A}_p \end{bmatrix} = \begin{bmatrix} \rho_p \\ 0_{p\times 1} \end{bmatrix}, \tag{11.18}$$

où la matrice $\mathcal{R}_p = \begin{bmatrix} R_0 & R_1 & \dots & \dots & R_p \\ R_1 & R_0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & R_0 & R_1 \\ R_p & \dots & \dots & R_1 & R_0 \end{bmatrix} \in \mathcal{M}_{p+1,p+1}(\mathbb{R}^+)$ est connue

et où le vecteur $\mathcal{A}_p = \begin{bmatrix} a_1^{(p)} \\ a_2^{(p)} \\ \vdots \\ a_p^{(p)} \end{bmatrix} \in \mathcal{M}_{p,1}(\mathbb{R})$ et le scalaire ρ_p sont inconnus.

Résolution

Étape d'initialisation : Pour $p = 0$, on trouve que $\rho_0 = R_0$ (et $\mathcal{A}_0$ est vide).

Étapes de la récursion : Soit $p \geq 1$ et supposons le problème $\mathcal{P}_{p-1}$ résolu. Alors, on a

$$\mathcal{R}_p \begin{bmatrix} 1 \\ \mathcal{A}_{p-1} \\ 0 \end{bmatrix} = \begin{bmatrix} \rho_{p-1} \\ 0_{(p-1)\times 1} \\ \delta_p \end{bmatrix}, \tag{11.19}$$

où les p premières lignes sont naturellement satisfaites d'après $\mathcal{P}_{p-1}$ et où la dernière ligne introduit un certain coefficient δ_p donné par

$$\delta_p = R_p + \begin{bmatrix} R_{p-1}, \ldots, R_1 \end{bmatrix} \mathcal{A}_{p-1}. \tag{11.20}$$

Multiplions (11.19) par J_{p+1} à gauche et substituons $\mathcal{R}_p$ par $J_{p+1}\,\mathcal{R}_p J_{p+1}$ (en accord avec la propriété 1). On obtient alors

$$(J_{p+1})^2\, \mathcal{R}_p\, J_{p+1} \begin{bmatrix} 1 \\ \mathcal{A}_{p-1} \\ 0 \end{bmatrix} = J_{p+1} \begin{bmatrix} \rho_{p-1} \\ 0_{(p-1)\times 1} \\ \delta_p \end{bmatrix}. \tag{11.21}$$

Puisque $(J_{p+1})^2$ vaut l'identité, on peut factoriser $\mathcal{R}_p$ à gauche dans la combinaison linéaire d'équations (11.19)$-k_p$(11.21). L'écriture détaillée de cette équation donne

$$\begin{bmatrix} R_0 & R_1 & \ldots & \ldots & R_{p-1} & R_p \\ R_1 & R_0 & R_1 & & & R_{p-1} \\ \vdots & R_1 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & R_1 & \vdots \\ R_{p-1} & & & R_1 & R_0 & R_1 \\ R_p & R_{p-1} & \ldots & \ldots & R_1 & R_0 \end{bmatrix} \left(\begin{bmatrix} 1 \\ a_1^{(p-1)} \\ \vdots \\ \vdots \\ a_{p-1}^{(p-1)} \\ 0 \end{bmatrix} - k_p \begin{bmatrix} 0 \\ a_{p-1}^{(p-1)} \\ \vdots \\ \vdots \\ a_1^{(p-1)} \\ 1 \end{bmatrix} \right) = \begin{bmatrix} \rho_{p-1} \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \delta_p \end{bmatrix} - k_p \begin{bmatrix} \delta_p \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \rho_{p-1} \end{bmatrix}.$$

Si on choisit k_p de sorte à annuler la dernière ligne, c'est-à-dire tel que

$$\delta_p - k_p\, \rho_{p-1} = 0, \tag{11.22}$$

on retrouve une équation de la forme (11.18) pour laquelle

$$\begin{bmatrix} 1 \\ \hline \mathcal{A}_p \end{bmatrix} = \begin{bmatrix} 1 \\ \hline \mathcal{A}_{p-1} \\ 0 \end{bmatrix} - k_p \begin{bmatrix} 0 \\ \hline J_p\, \mathcal{A}_{p-1} \\ 1 \end{bmatrix} \tag{11.23}$$

$$\rho_p = \rho_{p-1} - k_p\, \delta_p. \tag{11.24}$$

Ceci fournit la solution du problème $(\mathcal{P}_p)$.

La forme la plus compacte de l'algorithme de Durbin-Levinson est obtenue à partir des équations récurrentes fournies par : ⓐ l'équation (11.22) dans laquelle on a substitué δ_p par (11.20), puis ⓑ la partie basse (sous le trait) de l'équation (11.23), et enfin ⓒ l'équation (11.24) dans laquelle on a éliminé δ_p grâce à (11.22). Ceci conduit au résultat suivant.

Algorithme 1 (Durbin-Levinson)
Initialisation : $\rho_0 := R_0$, $\mathcal{A}_0 := []$.
Récurrence : Pour p allant de 1 à P exécuter :

ⓐ $k_p := (R_p + \left[R_{p-1}, \ldots, R_1\right] \mathcal{A}_{p-1})/\rho_{p-1}$,

ⓑ $\mathcal{A}_p := \begin{bmatrix} \mathcal{A}_{p-1} \\ 0 \end{bmatrix} - k_p \begin{bmatrix} J_p \mathcal{A}_{p-1} \\ 1 \end{bmatrix}$,

ⓒ $\rho_p := \big(1 - (k_p)^2\big)\rho_{p-1}$.

Action	Nombre d'opérations		
	+ ou −	×	/
$\rho_0 := R_0$, $\mathcal{A}_0 := [\,]$	0	0	0
$k_p := (R_{p+1} + \left[R_p, \ldots, R_1\right] \mathcal{A}_{p-1})/\rho_{p-1}$	p	p	1
$\mathcal{A}_p := \begin{bmatrix} \mathcal{A}_{p-1} \\ 0 \end{bmatrix} - k_p \begin{bmatrix} J_p \mathcal{A}_{p-1} \\ 1 \end{bmatrix}$	p	p	0
$\rho_p := \big(1 - (k_p)^2\big)\rho_{p-1}$	1	2	0
Total	$2p+1$	$2p+2$	1
Total pour $1 \leq p \leq P$	$P(P+2)$	$P(P+3)$	P
	$= 2\,P\,(P+3)$		

TAB. 11.1 – Détail du nombre d'opérations en virgule flottante pour l'algorithme 1 : la complexité algorithmique est de $2\,P\,(P+3) \equiv \mathcal{O}(P^2)$.

Quelques remarques et résultats connus On peut montrer que si $\mathcal{R}_P$ est inversible, alors $|k_p| < 1$ $(1 \leq p \leq P)$ de sorte que toutes les quantités sont calculables[22]. De plus, le filtre $H(z)$ peut aussi s'écrire directement à partir des coefficients k_p (plutôt que a_p) grâce à la représentation dite "en treillis". Cette représentation permet une simulation efficace dans le domaine temporel (cf. [18, 1] pour plus de détails). Dans cette structure, les paramètres k_p s'interprètent comme des coefficients de réflexion qu'on peut relier aux réflexions que subiraient à chaque jonction les ondes aller/retour voyageant dans une succession de tubes acoustiques droits (cf. figure 11.4 et les considérations acoustiques en page 14).

Dans l'algorithme, la quantité ρ_p joue le rôle de l'inconnue $N\sigma^2\,(= \rho_N)$ de (11.16), dans le cas d'un modèle d'ordre p. Sa décroissance avec p peut s'interpréter de la façon suivante. Si l'on fait croître l'ordre p du modèle, le filtre s'adapte de mieux en mieux aux données. Ainsi, l'excitation associée n'a plus à prendre en charge certaines variations des données (désormais expliquées par le filtre), ce qui réduit l'écart-type σ.

[22] On trouvera une étude très précise du conditionnement numérique de l'algorithme dans [14].

11.3.4 Tests et pré-accentuation

L'algorithme 1 qui résout (H7-H8) est testé sur les trames repérées sur la figure 11.5 par des lignes verticales (voyelles a, u, e des expériences (E1) et (E3), $T_{trame} = 70\,ms$) pour des paramètres standard de téléphonie ($F_e = 8\,kHz$, $K = 5$). Les résultats sont présentés en figure 11.6, (E1,v_n) et (E3,v_n).

Observations Les formants sont assez bien capturés pour la voix chuchotée (E3) mais plutôt mal pour la note tenue (E1). De plus, pour une même voyelle de (E1) et (E3), les correspondances des formants sont globalement décevantes. Puisque la qualité de la source a tant d'impact, il faut raffiner (H1). Quelle modification simple pourrait apporter une amélioration ?

Dans (H1), le signal source a été supposé de moyenne nulle et ne privilégiant aucune fréquence (on parle souvent de bruit à spectre plat ou de bruit blanc par analogie avec la couleur et le spectre des ondes électromagnétiques). Or, d'une part, on constate que les spectres ne sont pas d'énergie nulle à la fréquence nulle. D'autre part, outre les résonances formantiques, leur énergie a systématiquement tendance à décroître avec la fréquence.

En dessous de $4kHz$, la pression acoustique rayonnée en champ lointain est quasi-proportionnelle à la dérivée temporelle du débit acoustique aux lèvres : le signal subit un gain de 6 décibels par octave. Par ailleurs, le débit glottique (signal source) à un comportement très passe-bas, d'autant plus pour les sons voisés. Le bilan, visible en figure 11.6 (E1,v_n)-(E3,v_n), reste un comportement passe-bas, à compenser pour être en bon accord avec $\mathcal{H}$.

Révision du modèle Un moyen simple de représenter le comportement *globalement passe-bas* du couple *signal de source et rayonnement* consiste à considérer finalement le signal ε_n comme un bruit blanc de type (H1) filtré par un passe-bas élémentaire (un intégrateur). Aussi, pour rendre efficace l'analyse par LPC, on compense cet intégrateur en appliquant un dérivateur (numériquement, une différence $D(z) = 1 - z^{-1}$) au signal[23] v_n :

(**H9**) ***Pour rendre efficace l'analyse par LPC sur le son vocal, on l'applique sur le signal*** $\delta v_n = v_n - v_{n-1}$ ***plutôt que*** v_n***. Le filtre dérivateur de fonction de transfert*** $D(z)$ ***est appelé*** **filtre de pré-accentuation.**

L'application du filtre de pré-accentuation conduit aux résultats (E3,δv_n) et (E1,δv_n) de la figure 11.6 qui sont en effet meilleurs que (E1,v_n) et (E3,v_n). Ceci est confirmé par l'analyse par LPC complète du signal vocal représentée en figure 11.7. On y retrouve bien les formants, leurs trajectoires temporelles et les fameux motifs composés de 4 formants remarqués dans l'observation 1 (page 11) et la figure 11.5. La similarité des enchaînements de (E1) et de (E3) est également bien capturée par la méthode.

[23] La commutation du filtre $H(z)$ de (H2) et de $D(z)$ est possible car, sur chaque trame, le filtre $H(z)$ est invariant dans le temps.

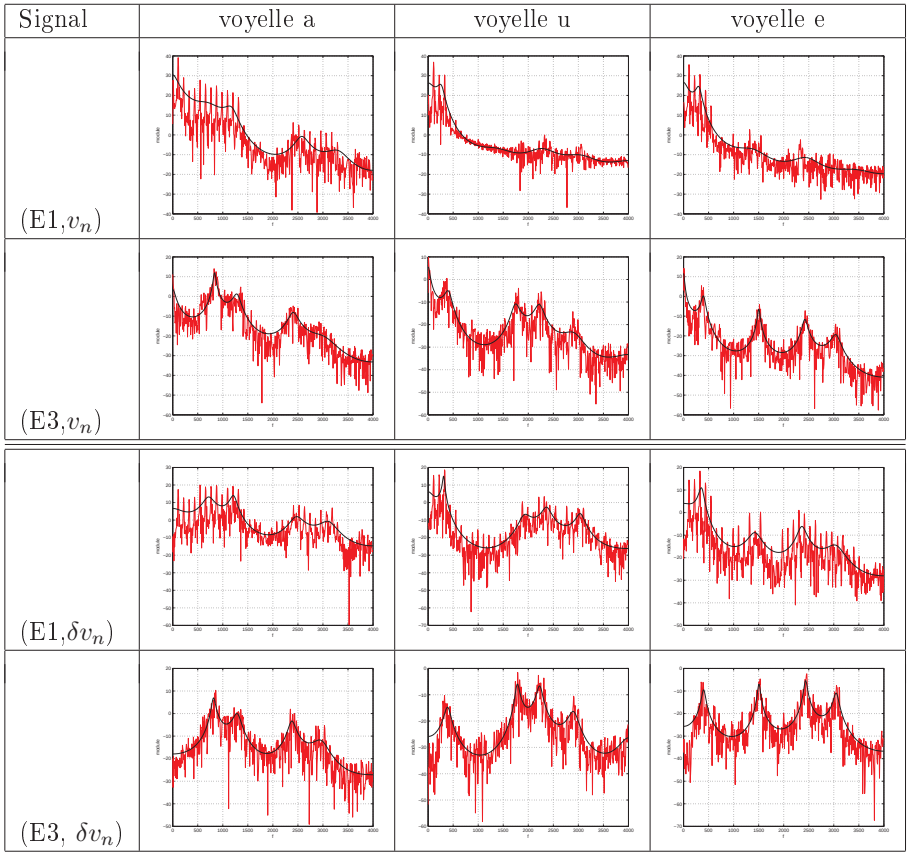

FIG. 11.6 – Spectre (–, en dB) et réponse fréquentielle des filtres (–, en dB) estimés par l'algorithme de Durbin-Levinson ($P=2K$, $K=5$) sur les signaux échantillonnés ($F_e = 8\,kHz$) de voix naturelle (v_n) et pré-accentuée (δv_n) pour les trames localisées en figure 11.5 par des lignes verticales (voyelles a, u, e des expériences (E1) et (E3), $T_{trame}=70\,ms$).

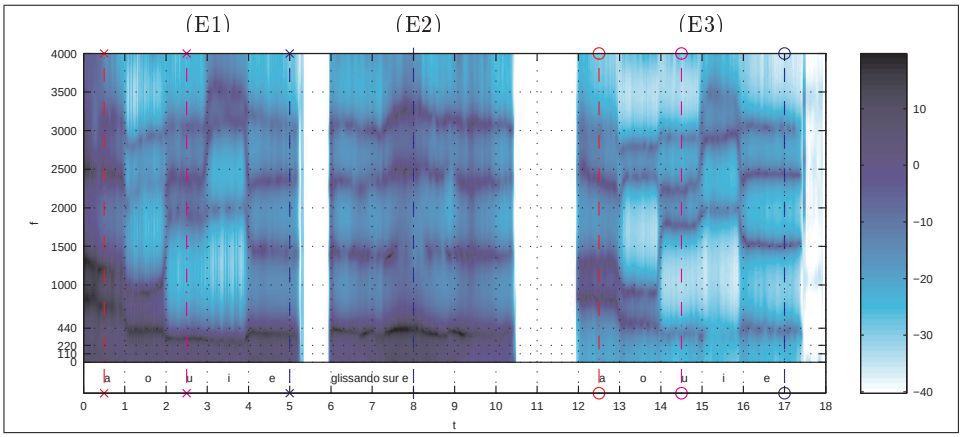

FIG. 11.7 – Réponses fréquentielles des filtres (niveau de couleur en dB) en fonction du temps de début de trame (signal δv_n, $T_{trame} = 70\,ms$, $F_e=8\,kHz$, $K=4$).

11.4 Travail pratique : construction du Vocoder

Vous voici maintenant maître du jeu. Le matériau fourni pour construire votre vocoder par LPC est le suivant :

a. le synopsis décrit en partie 11.4.1,
b. le canevas de code *Scilab* fournit en partie 11.4.2,
c. l'algorithme 1 détaillé page 24.

11.4.1 Synopsis

Les éléments décrits précédemment ont expliqué et résolu le problème de la séparation *excitation/conduit vocal*. La figure 11.3 rappelle le principe d'*analyse/synthèse* du vocoder. En pratique, pour avoir un résultat agréable à l'écoute, il reste deux points auxquels penser :

- à chaque trame de signal musical traité, il faut appliquer le filtre estimé mais aussi une amplification de gain σ, piloté par la source (un silence sur le signal de voix doit rester un silence sur le signal vocodé) ;
- gommer la rugosité sonore qui apparaîtrait si l'on se contentait de *juxtaposer* les "trames vocodées".

Pourquoi ce dernier point ? Les filtres estimés d'une trame à une autre sont distincts (même s'il sont proches). Ainsi, ces différences seront à l'origine de sauts à chaque raccord de trames. Ces artefacts apparaîtront avec la période du pas d'avancement de l'analyse et l'oreille l'entendra immédiatement. Une solution basique consiste à "gommer" ces artefacts en effectuant un fondu enchaîné à chaque raccord sur de courtes zones de recouvrement, comme cela est récapitulé dans la figure 11.8.

11.4.2 Code à compléter

Voici un canevas de code *Scilab* à partir duquel vous pourrez construire votre vocoder par LPC. En fait, la plupart des fonctions (y compris le calcul de l'auto-corrélation et l'algorithme de Durbin-Levinson) sont disponibles sous *Scilab*. Néanmoins, il pourra être intéressant de coder ces courtes fonctions (nommée ci-dessous, `AutoCorrel` et `Levinson`) vous-même et d'apprécier leur fonctionnement : vous pourrez en profiter pour comparer les résultats des fonctions *natives* aux vôtres afin de valider votre code et aussi comparer les temps d'exécution.

Le fichier `ProgVocoderLPC.sci` est le programme à appeler : il charge, prépare, vocode, normalise et sauve les sons au format `wav`. Les fonctions à coder par vos soins sont rassemblées dans le fichier `FonctionsVocoderLPC.sci`. Pour faciliter la lecture du code, les vecteurs sont notés avec le suffixe `_v` et les matrices (ou tableaux) avec le suffixe `_m`. Un exemple de programme d'exécution qui appelle ces fonctions est fourni dans le fichier `ProgVocoderLPC.sci`.

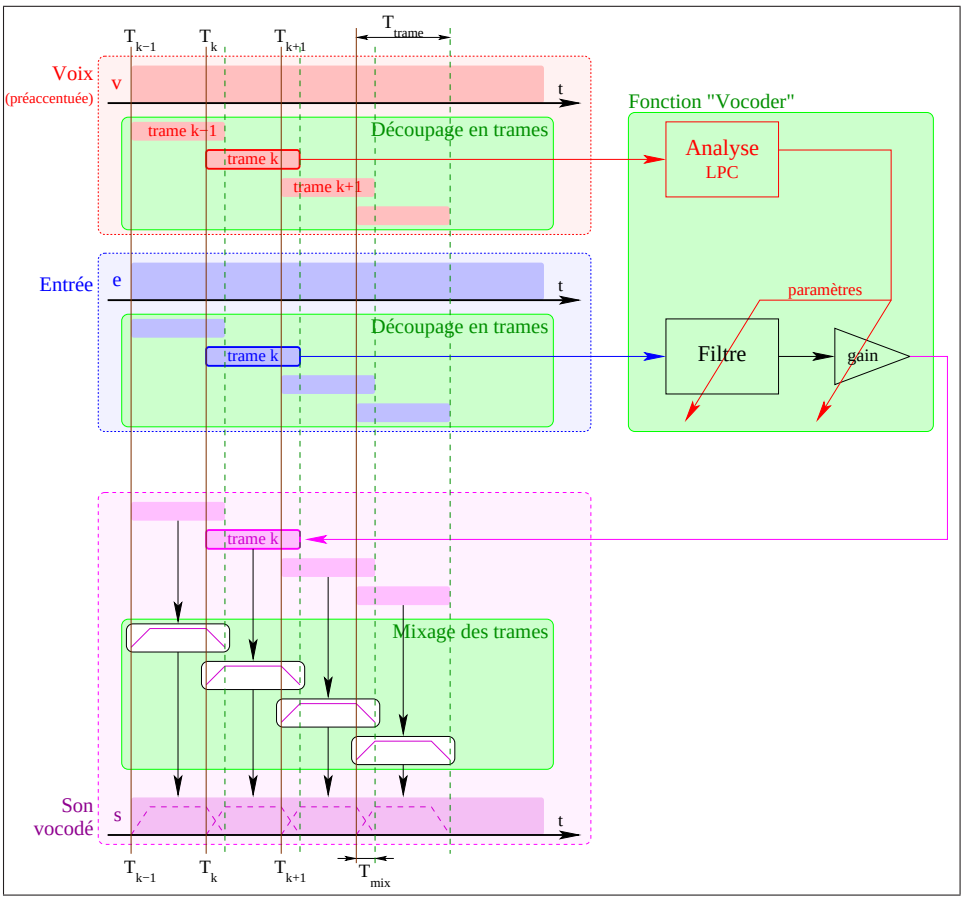

FIG. 11.8 – Synopsis du vocoder. Les fondus enchaînés sont construits à partir de gains appliqués à chaque trame : rampe linéaire croissante (de 0 à 1 au début d'une nouvelle trame), puis gain unitaire, puis gain décroissant (de 0 à 1 en fin de trame) de sorte que la somme des enveloppes de gain donne la constante 1. *Rq : Il faut veiller à la bonne synchronisation des signaux dans les zones de recouvrement.*

Fichier *FonctionsVocoderLPC.sci*

```
function [T_m]=DecoupeEnTrames(Sig_v,Fe,TpsTrame,TpsMix)

// A REMPLIR (la colonne i de la matrice T_m sera la ieme trame; pour eviter
// les boucles, construire un tableau d'indices Ind_m puis T_m(:)=Sig_v(Ind_m);

endfunction;


function [Sig_v]=RemixTrames(T_m,Fe,TpsTrame,TpsMix);

// A REMPLIR

endfunction;


function [A_v,sigma2,K_v]=Levinson(R_v);
// [A_v,sigma2,K_v]=Levinson(R);
// A_v=[1,a1,...,a_ordre].' est le vecteur des coefficients du filtre
// Rq: 1. ordre=length(R)-1
//     2. sigma2 est la variance de la variable aleatoire e_n
//     3. K_v est le vecteur des coefficients de reflexion

// A REMPLIR

endfunction;


function [R_v]=Autocorrel(V_v,n)
// V_v: vecteur signal
// n  : nombre de coefficients d'autocorrelation
// R_v: coefficients de l'autocorrelation

// A REMPLIR (pour eviter les boucles, utiliser indices et produit matriciel)

endfunction;


function SonVocoder_v=VocoderLPC(Voix_v,Musique_v,Fs,OrdreLPC,TpsTrame,TpsMix)

// 1. Preaccentuation du signal de voix
Voix_v=diff(Voix_v);

// 2. Formattage en vecteur colonne avec longueurs identiques
N = min([length(Voix_v),length(Musique_v)]);
Voix_v    = Voix_v(1:N);      Voix_v    = Voix_v(:);
Musique_v = Musique_v(1:N);   Musique_v = Musique_v(:);

// 2. Decoupage en Trames: TVoix_m et TMusique_m
// A REMPLIR

// 3. Construction des trames de sortie (analyse LPC et filtrage
//    pour chaque trame)
[NbEchTrame,NbTrames] = size(TVoix_m);
TSortie_m = zeros(NbEchTrame,NbTrames);
for n=1:NbTrames,
// A REMPLIR
end;

// 4. Remixage des trames stockees dans TSortie_m
// A REMPLIR

endfunction;
```

Fichier *ProgVocoderLPC.sci*

```
stacksize(4000000);
exec('FonctionsVocoderLPC.sci');

// 1. Sons et parametres
DirSon         = '';
FichierVoix    = 'voix.wav';
FichierMusique = 'musique.wav';

Fe       = 16000;   // frequence d'echantillonage consideree
OrdreLPC = 2*8;     // choix de 8 formants maximum pour Fe=16kHz
TpsTrame = 20e-3;   // en seconde
TpsMix   = 5e-3;


// 2. Chargement des sons monophoniques (pour les sons stéréophoniques,
//    selectionner une seule voie ou la moyenne des deux voies)
[Voix_v,Info]    = loadwave(DirSon+FichierVoix);
FVoix          = Info(3);
[Musique_v,Info] = loadwave(DirSon+FichierMusique);
FMusique       = Info(3);

// 3. Re-echantillonnage
Voix_v    = intdec(Voix_v(:,1),Fe/FVoix);
Musique_v = intdec(Musique_v(:,1),Fe/FMusique);


// 4. Appel au vocoder
SonVocoder_v = VocoderLPC(Voix_v,Musique_v,Fe,OrdreLPC,TpsTrame,TpsMix);

// 5. Normalisation pour le format wav (plage=[-1,+1]) et sauvegarde
SonVocoder_v = SonVocoder_v/max(abs(SonVocoder_v));
savewave(DirSon+'SonVocoder.wav',SonVocoder_v,Fe);
```

11.4.3 Exemple

Un son résultant du vocoder par LPC tel qu'il est présenté en figure 11.8 a été calculé. La voix prononce la première phrase de ce chapitre : *Peut-on faire "parler la musique" ou, plus généralement, un son ?* Le son correspondant est disponible sur le CD (plage 24). La musique est le début du premier mouvement de *La passion selon St Matthieu* de J. S. Bach (CD, plage 25). Le résultat est présenté en figure 11.9 et sur le CD (plages 26 et 27).

Sur cette figure, il apparaît que la méthode par LPC a extrait les formants (en ⓑ) portés par la structure fine du spectre (en ⓐ) du signal vocal. Ce "filtrage formantique" ⓑ a été appliqué au signal musical ⓒ pour fournir le son vocodé ⓓ. Sur le CD (plage 28), vous trouverez le résultat obtenu en choisissant un bruit blanc gaussien plutôt qu'un son musical.

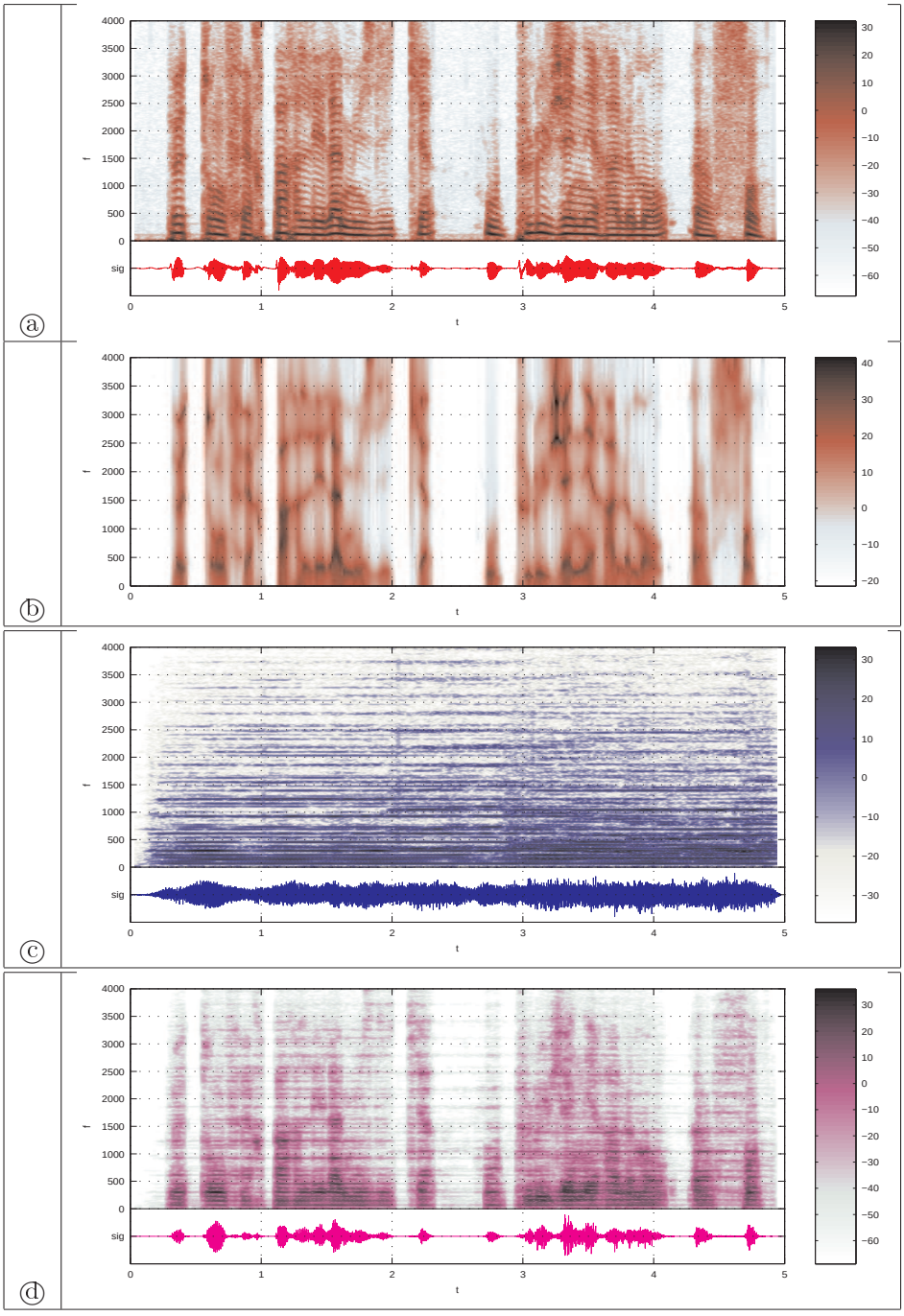

FIG. 11.9 – Spectrogrammes et signaux temporels de la voix ⓐ, de la musique(La passion selon St Matthieu, J. S. Bach) ⓒ, du son vocodé ⓓ et, en ⓑ, représentation des filtres estimés par la méthode par LPC avec pré-accentuation du signal vocal (niveau de couleur en dB). Les paramètres des spectrogrammes sont les mêmes qu'en figure 11.5. Les paramètres d'analyse-synthèse pour le vocoder sont : $T_{trame} = 20\,ms$, $T_{mix} = 5\,ms$, $F_e = 8\,kHz$, $K = 5$ (paramètre souvent pris en téléphonie).

11.5 Pour aller plus loin

Dans ce chapitre, un présentation intuitive d'un *modèle de production de la voix* a été présenté, qui a permis de bâtir la méthode par LPC pour construire un *vocoder musical*. Ces résultats a été présentés et obtenus en se limitant à un niveau de connaissance sur la production vocale volontairement restreint.

Bien d'autres choix techniques sont possibles et ont été étudiés (représentation du filtre différente de (11.1), autres méthodes d'estimation, etc.). Des niveaux de description plus fins (modèle de signal glottique, modèle aéro-acoustique de la voix, etc.) peuvent aussi apporter des résultats plus précis et réalistes. Les travaux existants sur ces sujets sont beaucoup trop nombreux pour les citer ou même en faire un abrégé.

Outre les références bibliographiques qui suivent, terminons donc en indiquant quelques revues et conférences scientifiques dans lesquelles vous pourrez trouver bon nombre de ces travaux :

- Approches "traitement de signal" :
  - IEEE Transactions on Audio, Speech and Language Processing (IEEE TASLP),
  - IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP)
- Approches "acoustiques" :
  - Acta Acustica (S. Hirzel Verlag),
  - Journal of the Acoustical Society of America (JASA, American Institue of Physics),
  - Journal of Sound and Vibration (JSV, Elsevier).
- Travaux spécialement dédiés à la voix :
  - Interspeech
  - Pan-European Voice Conference (PEVOC)

Bibliographie

[1] M. Benidir. *Quelques représentations d'un polynôme et leurs applications en traitement du signal*, Traitement du Signal, 15(6), 1998.

[2] P. Boersma and D. Weenink. *Praat : doing phonetics by computer [Computer program]*, Version 5.2.16, retrieved 20 february 2011 from `http ://www.praat.org/`. *Ce logiciel libre est bien documenté. Il inclut une boîte à outils d'analyse-transformation dont la méthode par LPC.*

[3] D. Butler. *You'll find out*, Film, 1940. *On peut apprécier le sonovox dans une partie avec Kay Kyser et son orchestre, en "faux direct" monophonique : bonne découverte !.*

[4] H. Coker C. B. Denes P. and N. Pinson E. *Speech/synthesis : an experiment in electronic speech production*, Number 3 in Bell system science experiment. Bell Telephone Laboratories, Baltimore, Md., Waverly Press, Inc. edition, 1963.

Une initiative des Bell Labs était de proposer en kit des expériences issues de recherches approfondies. Ainsi, en 1963, contre quelques dizaines de dollars, vous receviez par la poste un colis incluant livret explicatif, schémas électroniques et composants à enficher sur une plaque de carton préparée pour faire vos premières synthèses de voix. Mots clef (web) : Bell Labs Science Experiment Kits *(no. 3).*

[5] Calliope. *La Parole et son traitement automatique*, Masson, 1989.
(Le livre français de référence, encore à ce jour).

[6] H. W. Dudley. *Signal transmission*, United States Patent Office, No.2,151,091 (Application, October 30, 1935, Serial No.47, 393), Patented March 21, 1939.
(Ce brevet qui décrit le vocoder construit en 1935 par H. Dudley est accessible sur le web).

[7] H. W. Dudley and H. Tarnoczy T.*The speaking machine of Wolfgang von Kempelen*, The Journal of the Acoustical Society of America, 1950.

[8] G. Dournon-Taurelle et J. Wright. *Les Guimbardes du Musée de l'Homme*, Muséum national d'histoire naturelle, Institut d'ethnologie, 1978. *(Voir aussi "M. Wright, The Search for the Origins of the Jew's*

Harp, Oxford, England", les sites web associés mais aussi un instrument du même type "musical bow" : un sujet très riche).

[9] R. M. Gray. *Linear Predictive Coding and the Internet Protocol : A survey of LPC and a History of Realtime Digital Speech on Packet Networks*, Now Publishers, 2010.
Ce livre est rédigé par l'un des pionniers de la méthode par LPC. La première partie rassemble et explique les nuances entre plusieurs approches optimales et plusieurs choix possibles (souvenez-vous de nos hypothèses (H4) et (H5) par exemple !), qui conduisent à des méthodes de codage par prédiction linéaire. La seconde partie retrace l'épopée des chercheurs, des ingénieurs, de leurs inventions et réalisations.

Outre les très sérieuses applications pour les télécomunications, on y apprend (p.113-114) que le codage par prédiction linéaire *fut incorporé dans un circuit intégré utilisé pour l'un des premiers* jeux parlants pour enfant*, "la dictée magique" (Speak and Spell), dont on retrouve un exemplaire dans le célèbre film E.T. de S. Spielberg.*
Une version électronique libre de ce livre est disponible sur la page de l'auteur `http ://ee.stanford.edu/ gray/lpcip.pdf` .

[10] N. Henrich. *Étude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception*, Thèse, Université Paris 6, 2001.
(Cette thèse se focalise sur le signal source dont l'étude a justement été laissée de côté dans ce chapitre. On y trouve entre autres des renseignements sur les différents mécanismes de phonation dont les plus connus correspondent à la "voix de tête" et la "voix de poitrine").

[11] D. Klatt. *Review of text-to-speech conversion for English*, J. of the Acoustical Society of America, 82(3), 1987.

[12] F. Le Huche. *Réhabilitation vocale après laryngectomie totale*, Masson, 1997.

[13] E. Leipp. *Un "vocoder" mécanique : la guimbarde*, Annales des télécommunications, 18(5-6), mai-juin 1963.

[14] E. Kazamarande et P. Comon. *Stabilité numérique de l'algorithme de levinson*, Modélisation mathématique et analyse numérique (MAN), 29(2), 1995, pp. 123–170.

[15] W. von Kempelen. *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine, et Le Méchanisme de la parole, suivi de la description d'une machine parlante*, J.V. Degen. Vienna. 1791.

[16] J.-S. Liénard. *Reconstruction de la machine parlante de Kempelen*, Proceedings of the 4th International Congress of Acoustics, Budapest 1967.

Remarque : Un colloque en mémoire du 200ième anniversaire de la disparition de Farkas Kempelen (1734-1804) a eu lieu également à Budapest en mars 2004. On peut trouver le programme sur le web (International Workshop in Phonetics dedicated to the memory of Farkas Kempelen, Budapest, March 11-13, 2004).

[17] J.-S. Liénard. *Les processus de la communication parlée ; introduction à l'analyse et à la synthèse de la parole*, 189 p., Masson, Paris, 1977.

[18] J. Laroche. *Traitement des Signaux Audio-Fréquences*, TELECOM Paris, 1995.

[19] J. D. Markel and Jr. A. H. Gray. *Linear Prediction of Speech*, Springer Verlag, Berlin, 1976.

[20] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*, Prentice-Hall, Upper Saddle River, New Jersey, 1999.

[21] W. Pereira. *Modifying LPC Parameter Dynamics to Improve Speech Coder Efficiency*, PhD thesis, McGill University, Montréal, Canada, 2001.

[22] B. Picinbono. *Théorie des signaux et des systèmes avec problèmes résolus*, Collection Pédagogique de Télécommunication. Dunod, Paris, 1989.

[23] H. Reinhard. *Éléments de mathématiques du signal (tome 1 - Signaux déterministes)*, Dunod, 1995.

[24] M. R. Schroeder and B. S. Atal. *Code-excited linear prediction (celp) : high-quality speech at very low bit rates*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 10, 1985, pp. 937–940.

[25] J. Smith. *Tearing speech to pieces : Voice technologies of the 1940s*, Music, Sound, and the Moving Image, Liverpool University Press, 2(2), 2008, pp. 183–206.