Quelques idées directrices en reconnaissance automatique de la parole

J.-S. LIENARD *, M. CASTELLENGO **, E. LEIPP **, M. MLOUKA *, G. RENARD *, J. SAPALY **, D. TEIL *

Automatic recognition of speech is studied on the same bases used in synthesis by the same coworkers. The main difficulties of the problem are briefly shown ,and several experiments simulated on a large computer are described. A speech analysis procedure preserving the essential part of its information is defined, followed by some pretreatments making the analysis available for any speaker. Two different approaches of the recognition process are then proposed. The first one consists in isolating phonetic items (phonatoms) from continuous speech, before identifying them. The second one deals with isolated words; these are considered as a whole by an algorithm brings to a balance time local differences existing between two utterances of the same word.

Malgré le pessimisme de certains spécialistes éminents il faut reconnaître que la reconnaissance automatique de la parole est plus que jamais une question d'actualité. L'intérêt porté à ce problème par de nombreux chercheurs est en partie justifié par son enjeu économique et scientifique. En effet, de nombreux champs d'application sont actuellement ouverts : dialogue homme-machine, commande vocale, compression des bandes téléphoniques, aide aux handicapés, recherche linguistique, etc... Mais surtout l'approche du problème a profondément changé depuis quelques années, et l'on peut penser que la déception suscitée par de nombreuses expérimentations — pourtant très généralement présentées comme des réussites — est due en réalité à la sous-estimation des difficultés à surmonter, lesquelles sont essentiellement d'ordre psycho-physiologique et non d'ordre physique.

Le regain d'intérêt pour la reconnaissance automatique de la parole est sensible à l'étranger comme en France, où plusieurs équipes poursuivent ou abordent des études dans ce domaine. Certaines d'entre elles bénéficient de l'appui de la Direction des Recherches et Moyens d'Essais ou du Comité de Recherche en Informatique. Les motivations, les hypothèses et les outils de recherche sont propres à chaque équipe, et témoignent d'une grande diversité. Les appareillages utilisés, par exemple, peuvent être essentiellement analogiques [1, 2] (1), essentiellement numériques [3], mais se composent le plus souvent de la partie analyse d'un vocoder [4, 5, 6, 7] ou d'un banc de filtres spécialement adapté au problème [8, 9], qui permet de n'introduire en machine qu'un signal déjà dépouillé d'une partie de sa redondance. La plupart des chercheurs s'intéressent aux phonèmes (voyelles et consonnes), mais certains se préoccupent aussi des transitions entre phonèmes [8], ou considèrent des unités phonétiques voisines de la syllabe [5, 7], voire des mots entiers, séparés les uns des autres par des silences [4]. Les opinions diffèrent également au sujet de l'apprentissage, que l'on cherche à éviter [1, 2] ou au contraire à rationaliser au moyen de matrices [8] ou d'opérateurs algorithmiques [6]. L'utilisation des contraintes d'ordre syntaxique et sémantique a été étudiée dans le cadre réduit d'un langage de programmation [7], et préoccupe actuellement la majorité des chercheurs.

Nous essaierons, dans les lignes qui suivent, de poser le problème de la reconnaissance vocale automatique sur des bases qui nous sont propres, et de citer à l'appui de ces idées les expérimentations que nous avons pu effectuer dans le cadre du contrat nº 71/13 avec le Comité de Recherche en Informatique. Les études que nous avons menées au sujet de la synthèse, qui ont naturellement contribué à orienter nos recherches en reconnaissance automatique, ont déjà fait l'objet de plusieurs publications [10, 12, 13] notamment dans cette même Revue [11], et nous invitons le lecteur à s'y reporter s'il désire de plus amples détails sur l'analyse spectrographique de la parole ou le fonctionnement du synthétiseur Icophone. L'ensemble du travail effectué sur la reconnaissance a également été exposé par ailleurs de manière plus approfondie [15].

* Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, C.N.R.S., Orsay (Pr Malavard).

** Laboratoires d'Acoustique et d'Electronique, Institut de Mécanique Théorique et Appliquée, Université Paris VI (Pr Siestrunck).

(1) Les chiffres entre crochets renvoient à la bibliographie *in fine*.

1. PRINCIPALES DIFFICULTÉS DE LA RECONNAISSANCE AUTOMATIQUE

1.1. Difficultés d'ordre sémantique

La communication entre individus n'obéit pas au simple schéma de la théorie de l'information, composé d'un émetteur, d'un canal et d'un récepteur. D'une part, il existe toujours simultanément plusieurs canaux sensoriels (canal auditif, visuel, tactile, etc...); d'autre part, le processus de compréhension est actif. Le récepteur ne se contente pas de comparer passivement les informations reçues à celles qui sont stockées dans sa mémoire, mais prend l'initiative d'un certain comportement vis-à-vis de l'information à venir, en fonction de son intelligence, de son expérience, de sa connaissance du langage utilisé, et du contexte à l'instant présent. Ces facteurs, extrêmement difficiles à définir objectivement et variables au rythme de la vie même, font qu'il est illusoire, actuellement, de vouloir reproduire artificiellement le processus de la compréhension humaine.

Cependant le message parlé naturel est élaboré en fonction de récepteurs intelligents, au sens humain du terme. Par exemple l'analyse acoustique d'une phrase de la conversation courante révèle très souvent l'absence ou la déformation d'une partie des sons qui devraient y figurer, et que l'auditeur a la sensation d'entendre. Le locuteur articule d'autant moins clairement, et l'auditeur comprend d'autant plus facilement, que les facteurs précédents (canaux sensoriels multiples, intelligence, contexte, etc...) sont plus importants.

Il ne faut donc pas demander à un système de reconnaisance automatique plus qu'il ne peut donner : si l'on cherche, par exemple, à reconnaître les sons élémentaires de la parole (phonèmes, phonatomes) sans s'intéresser aux niveaux syntaxique et sémantique, il est nécessaire soit d'imposer une prononciation particulière, soit de tolérer un taux d'erreurs phonétiques variant de 0 à 100 % selon le locuteur et l'instant considéré.

1.2. Difficultés tenant à la continuité du message parlé

L'auditeur comprenant un message parlé a la sensation d'entendre des mots séparés. La réalité physique est tout autre : à l'intérieur d'un même « groupe de souffle » limité par deux ponctuations ou respirations, les seules interruptions sont constituées des courts silences précédant les explosions de certaines consonnes (p, t, k). Il n'existe pas, sauf artefact, de séparation physique entre les mots.

C'est pourquoi la reconnaissance vocale, depuis ses débuts en 1950, connaît deux écoles. L'une cherche à isoler puis identifier des éléments (ou segments) tels que phonèmes, phonatomes, syllabes, dans le flot continu de parole. L'autre simplifie le problème au départ en ne considérant que des mots prononcés isolément; le terme de « mot » est ici considéré dans un sens très large : il peut s'agir de plusieurs mots prononcés continûment ou d'une phrase complète, considérée comme un tout.

Les méthodes analytiques dirigées vers la reconnaissance en temps réel ou légèrement différé, utilisent surtout des techniques analogiques et des volumes réduits de calcul ou de mémoire. Elles sont à l'origine de nombreuses déceptions, car nous avons vu plus haut que le message acoustique ne contient généralement pas toute l'information phonétique de manière objective. La segmentation, en particulier, est une opération délicate, surtout si les éléments choisis sont les phonèmes.

Les méthodes globales, par contre, semblent plus prometteuses : la limitation du vocabulaire à quelques dizaines de mots fait que, généralement, chaque mot, même imparfaitement prononcé, contient suffisamment d'information pour être identifié. Mais tous les problèmes ne sont pas résolus pour autant : il suffit d'augmenter la taille du vocabulaire ou de changer de locuteur pour retrouver les difficultés d'ordre sémantique. L'identification des mots se fait nécessairement en différé; mais pour que le système soit susceptible d'application pratique, le délai requis ne doit pas dépasser quelques secondes avec un ordinateur de petite taille. Enfin les traitements destinés à normaliser les mots ou à en extraire les paramètres utiles avant l'identification ne sont pas moins épineux qu'en reconnaissance analytique.

1.3. Difficultés dues à la variabilité du message physique

Si l'on demande à un même locuteur de prononcer la même phrase de manière identique à quelques instants d'intervalle, on peut s'attendre à ce que le signal physique soit sensiblement identique dans les deux cas. Cela n'est jamais vérifié : l'analyse temps-fréquence révèle toujours des différences, qui peuvent être considérables, notamment dans la répartition des durées des sons composants, dans l'évolution du spectre de raies et dans l'intensité sonore. Quand aucune contrainte de prononciation n'est imposée, la variété des messages physiques devient très grande : le locuteur peut adopter divers types de voix (voix parlée ordinaire, voix chuchotée, chantée, criée, etc...), divers timbres, divers rythmes, etc... (fig. 1). Dans ce cas il est très difficile d'extraire automatiquement du signal les paramètres physiques caractérisant le sens du message.

D'un locuteur à l'autre on observe a fortiori ces mêmes différences, auxquelles s'ajoutent les différences d'ordre linguistique (accents régionaux, p. ex.), morphologiques (dimensions du conduit vocal variant selon l'âge, le sexe, etc...) et physiologiques (habitudes articulatoires).

Il est donc bien évident que le problème de la reconnaissance automatique, envisagé dans toute sa généralité, est d'une complexité inouïe. Les performances d'un système de reconnaissance dépendent essentiellement des contraintes imposées aux locuteurs, celles-ci étant rarement explicitées dans la littérature — tout simplement parce qu'elles sont difficiles à définir objectivement. Il s'ensuit que les taux de

réussite annoncée à la suite de chaque expérimentation doivent être accueillis avec le plus grand discernement, sous peine de s'exposer aux désillusions évoquées plus haut.

Nous allons maintenant envisager les différentes étapes d'un système de reconnaissance simulé sur ordinateur, à savoir l'analyse de la parole et sa représentation, sa normalisation en temps et en fréquence, et sa reconnaissance selon une optique analytique, puis selon une optique globale [15]. Nous ne nous attacherons pas à obtenir des taux de réussite élevés, mais à indiquer des voies de recherche possibles, et à justifier leur bien-fondé.

2. ANALYSE DE LA PAROLE

Dans nos études de synthèse, nous avons mis en évidence la notion de squelette informatif de la parole. Il s'agit d'une forme, ou structure perceptive, qui peut être schématisée dans le plan temps-fréquence (fig. 2), et qui représente le codage acoustique de l'ensemble des mouvements effectués par l'organe phonatoire du locuteur. L'information contenue dans ce schéma est beaucoup plus faible que l'information du signal entier, car le spectre de raies et les nuances d'intensité sont éliminés. La première étape de la reconnaissance automatique consiste donc à effectuer une analyse de la parole qui préserve le squelette informatif et élimine l'information superflue. Plusieurs points sont à considérer.

2.1. Régulation du niveau sonore

Le niveau sonore de la parole est une grandeur extrêmement variable : un même locuteur peut parler plus ou moins fort, s'approcher du microphone, tourner la tête, etc...; le phénomène se produit a fortiori d'un locuteur à l'autre, et les variations de niveau sonore atteignent rapidement 6, 12, voire 20 dB, ce qui correspond à des rapports de 2, 4 ou 10 en amplitude spectrale. Ces variations, à peine remarquées par l'oreille, sont particulièrement gênantes en reconnaissance automatique, car les configurations à comparer doivent être normalisées en niveau. Le problème est moins aigu si les configurations sont paramétrisées, p. ex. si l'on s'intéresse à l'évolution des fréquences de formants. Si l'on tient à conserver la globalité de la configuration temps-fréquence, on peut effectuer la normalisation dynamique soit de manière instantanée, soit sur l'ensemble de la configuration à étudier, ce qui ne peut se faire qu'en différé. Nous avons choisi la première méthode pour la reconnaissance analytique, afin de pouvoir la transposer éventuellement dans le domaine analogique.

La régulation dynamique doit, à notre sens, avoir un double but :

— elle doit maintenir à peu près constant le niveau sonore moyen, estimé sur une période d'au moins 200 ms en parole courante;

— elle doit accentuer certaines modifications spectrales significatives du message, comme les bruits d'impact rencontrés dans les plosives. Il est nécessaire pour cela de tenir compte du niveau sonore du message, non seulement à l'instant présent, mais aussi dans le passé immédiat. La correction tend par exemple à augmenter le niveau si le message est faible, mais ceci d'autant plus qu'il est faible depuis plus longtemps.

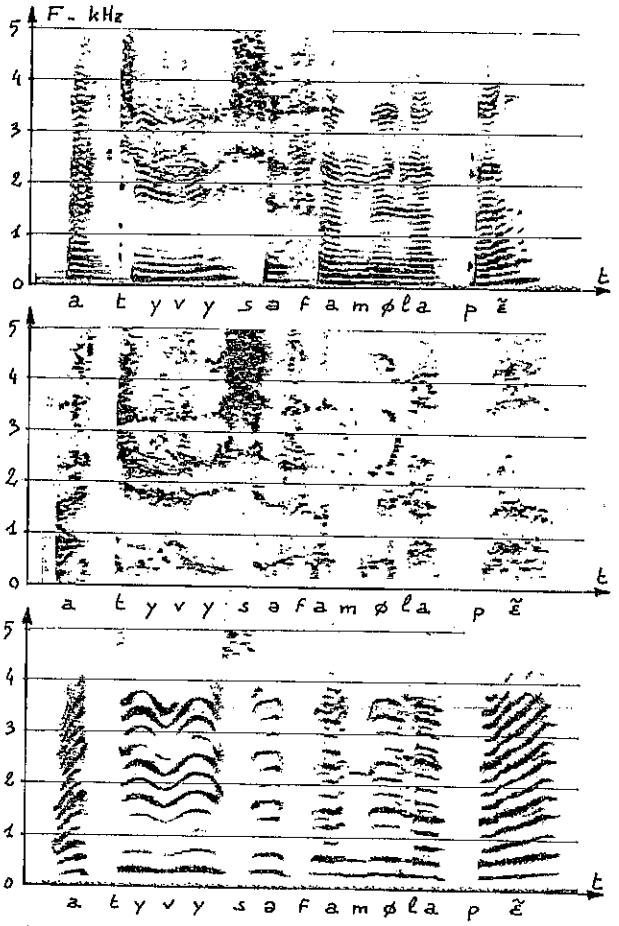


Fig. 1 — Divers sonagrammes, en bande étroite (Δ F = 45 Hz) de la phase « as-tu vu ce fameux lapin? », prononcée par un locuteur masculin en voix « normale », puis en voix chuchotée et en voix criée

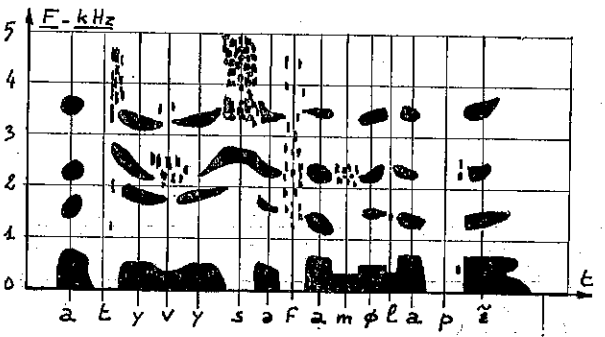


Fig. 2 — Squelette informatif de la phrase : « As-tu vu ce fameux lapin? », directement utilisable en synthèse

Il existe en électroacoustique des appareils appelés compresseurs de dynamique, qui effectuent analogiquement des traitements du même ordre. Cependant, leur destination étant différente, il convient d'examiner de près les possibilités de réglage qu'ils offrent, notamment en ce qui concerne les constantes de temps.

2.2. Analyse temps-fréquence

L'analyse fréquentielle de la parole est soumise à des impératifs contradictoires, qui trouvent leur origine dans certaines particularités de la perception auditive [15]. Sans développer ce chapitre on peut admettre en première approximation que la largeur du filtre d'analyse doit être de l'ordre de 300 Hz pour éliminer le spectre de raies; mais l'intervalle de fréquence entre deux filtres consécutifs doit être nettement inférieur si l'on veut pouvoir séparer les formants voisins dans certaines voyelles. La définition temporelle doit être de l'ordre de 10 ms pour préserver les transitoires rapides rencontrés dans les consonnes plosives. Quant à la bande passante de la parole, il convient de la limiter à environ 400 - 5 000 Hz.

2.3. Pondération fréquentielle

C'est un fait bien connu que l'énergie moyenne de la voix décroît de 6 à 10 dB par octave, au-delà d'une fréquence située entre 500 et 1 000 Hz, ceci avec des variations assez considérables d'un locuteur à l'autre. Il est donc nécessaire d'appliquer une pondération renforçant les valeurs spectrales situées au-dessus de 1 000 Hz. Il faut de plus tenir compte du maximum de sensibilité de l'oreille entre 1 000 et 3 000 Hz. Cette zone correspond au domaine d'évolution du 2e formant, qui est le plus significatif. Ces considérations nous amènent à utiliser une courbe de pondération moyenne renforçant considérablement les fréquences voisines de 1 500 Hz (fig. 3). En réalité, chaque locuteur possède un timbre particulier, pour un type de voix et un niveau sonore donnés, et la courbe adoptée devrait être ajustée dans chaque cas en fonction de ces facteurs.

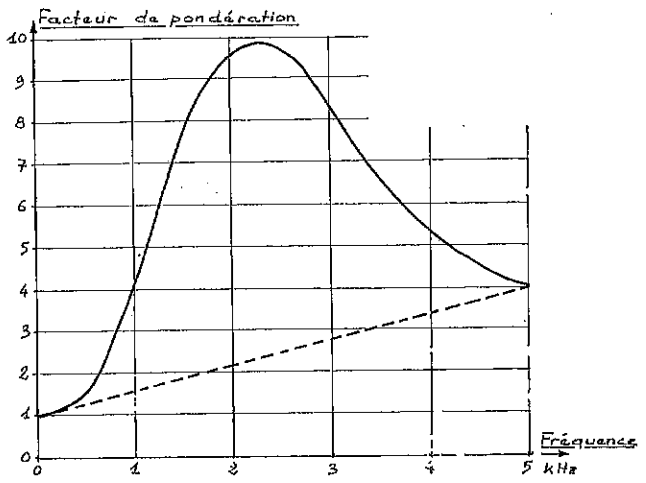


Fig. 3 — Pondération fréquentielle utilisée dans les essais de reconnaissance. Cette correction, appliquée à l'amplitude spectrale issue de l'analyse, résulte de deux sortes de considérations :

a) L'énergie moyenne de la voix décroît quand la fréquence augmente.

b) Les phénomènes importants ont lieu dans la zone sensible de l'oreille, centrée approximativement sur 2 000 Hz.

2.4. Simulation

Nous avons enregistré un certain nombre de phrases, mots et phonatomes, prononcés par 8 locuteurs différents, masculins et féminins. Nous n'avons pas cherché à faire des enregistrements de très haute qualité, de façon à rester dans les conditions usuelles de la transmission acoustique : le rapport signal-bruit est d'une trentaine de décibels, et le signal est accompagné d'une légère réverbération. Aucune contrainte de prononciation n'a été imposée aux locuteurs, si ce n'est la recommandation de parler à haute et intelligible voix.

Le codage analogique-numérique a été effectué par le Groupe de Recherche Ionosphérique du C.N.R.S.; le décodage, l'analyse et l'ensemble des traitements ultérieurs ont été programmés sur l'ordinateur IBM 360-75 du Centre de Calcul du C.N.R.S. (CIRCE), en PL/1 ou en Fortran selon les cas.

Le signal, numérisé à la cadence de 10 000 échantillons par seconde, a été analysé au moyen d'un programme de transformation rapide de Fourier dû à M. Delouis (CIRCE). Des prélèvements de 128 échantillons effectués toutes les 10 ms (donc avec un léger recouvrement) et pondérés selon la fenêtre de Hanning (arche de sinusoïde) ont fourni des spectres d'amplitude composés de 64 points répartis régulièrement entre 0 et 5 000 Hz. Un lissage a porté la largeur apparente d'analyse à environ 250 Hz, et la pondération fréquentielle de la figure 3 a été appliquée, ainsi qu'un processus de régulation de niveau, entièrement numérique, présentant les caractéristiques exposées en 2.1.

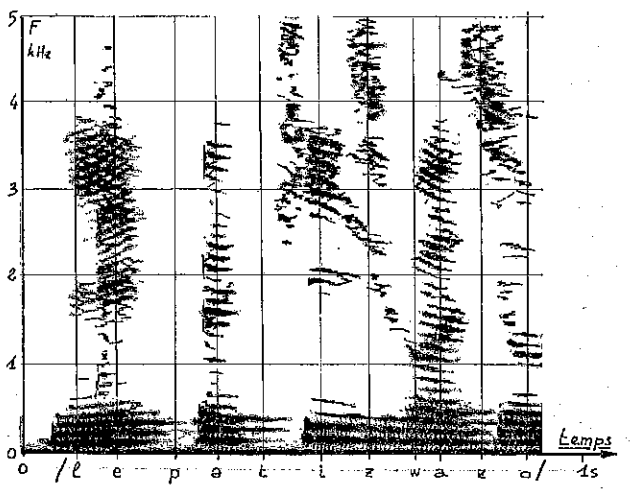


Fig. 4 — Sonagramme en bande étroite (largeur d'analyse 45 Hz) de la phrase : « Les petits oiseaux », prononcée par un locuteur masculin

Le résultat de l'analyse temps-fréquence est porté sur les figures 4 et 5, et permet de comparer le sonagramme ordinaire et le sonagramme numérique de la même séquence de parole. Pour l'édition sur imprimante, l'amplitude spectrale a été ramenée à 5 valeurs seulement, à savoir, dans l'ordre croissant : blanc, apostrophe, chiffre 1, chiffre 2, astérisque. Avec un minimum d'habitude de lecture des sonagrammes, on retrouve dans le document numérique toute l'information sémantique contenue dans le document analogique, ce qui constituait l'objectif de cette phase de la recherche.

3. NORMALISATION EN TEMPS ET EN FRÉQUENCE

La normalisation temporelle de la parole est nécessaire dans tout système de reconnaissance, même prévu pour un seul locuteur. Elle peut se déduire de l'étude de stationnarité (ou stabilité temporelle) du squelette informatif. La normalisation fréquentielle est une des clés de l'adaptation au locuteur et se rattache à la notion d'anamorphose.

3.1. Stationnarité du squelette informatif - Phonatome

Le squelette informatif est une succession d'états quasi-stables et d'états instables ou transitoires, qui reflètent acoustiquement les mouvements articulatoires. Pour apprécier cette stabilité il suffit de comparer le spectre à lui-même, avec un décalage de l'ordre de 30 ms (parole courante). Si le squelette informatif reste identique, la stabilité est maximale; elle passe par des minimums lors des transitions. C'est à dessein que nous n'employons pas ici le terme d'autocorrélation, qui possède une définition mathématique très précise. Mais il s'agit du même concept.

Nous appelons phonatome la portion du squelette informatif comprise entre deux maximums de stabilité. Chaque phonatome est l'image acoustique d'un mouvement élémentaire du conduit vocal; nous avons montré en synthèse l'utilité de cet élément phonétique, qui n'a que des relations très lâches avec la notion abstraite de phonème.

3.2. Taux de coïncidence

Pour comparer entre eux deux spectres ou deux configurations multidimensionnelles, nous avons mis au point un critère particulier de distance, appelé taux de coïncidence. C'est le rapport τ de l'intersection $A \cap B$ de deux ensembles A et B à leur réunion $A \cup B$, à une constante K près (fig. 6) :

$$\tau = \frac{K + A \cap B}{K + A \cup B}.$$

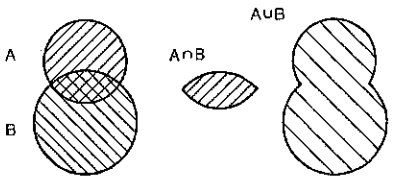


Fig. 6 — Intersection et union de deux ensembles, utilisées dans le calcul du taux de coïncidence

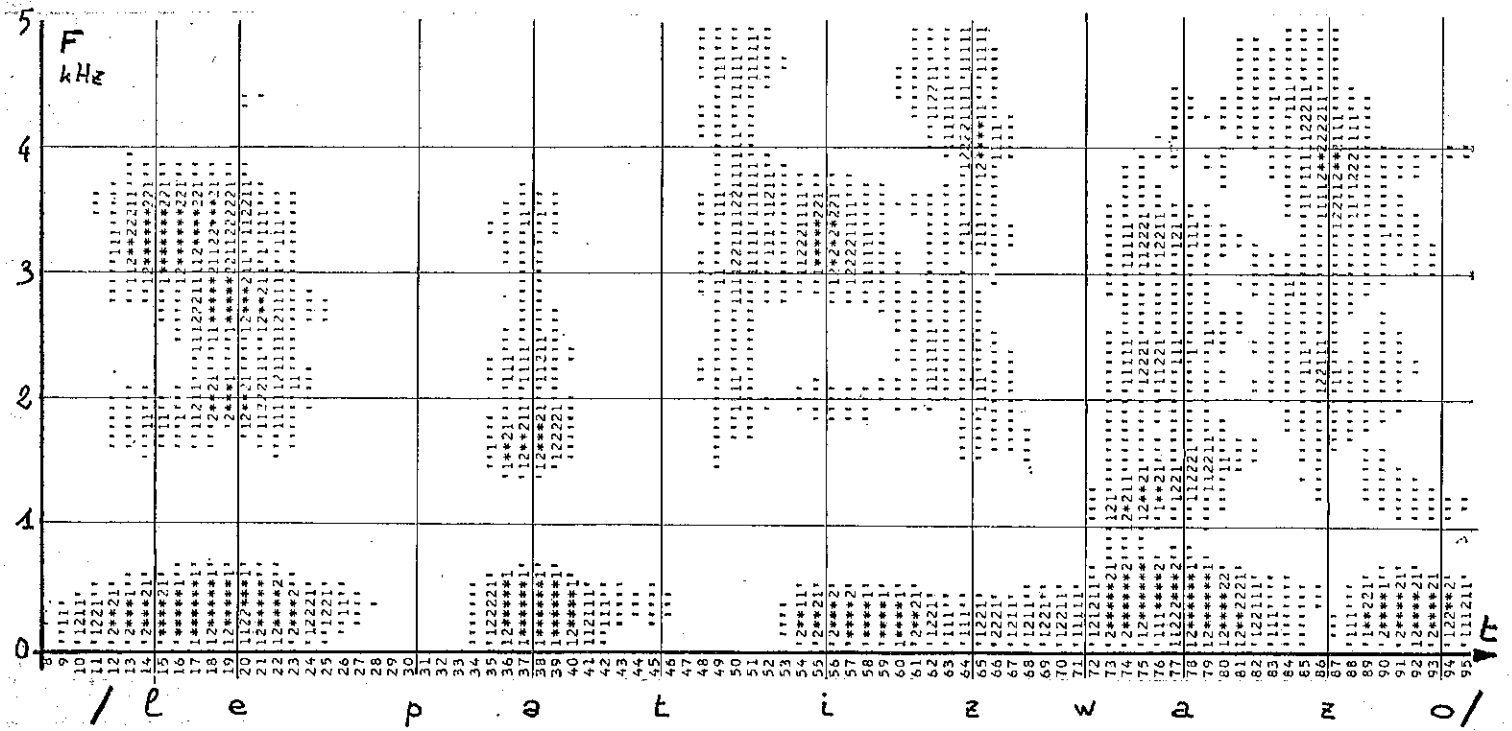


Fig. 5 — Sonagramme numérique de la même séquence, obtenu par transformation rapide de Fourier. Après pondération fréquentielle et lissage, la largeur apparente d'analyse est de l'ordre de 250 Hz

On peut aisément montrer que si a_{ij} et b_{ij} sont les éléments de A et B (configurations à deux dimensions, par exemple) le taux de coïncidence s'exprime par

$$\tau = \frac{K + \sum_i \sum_j \min(a_{ij}, b_{ij})}{K + \sum_i \sum_j \max(a_{ij}, b_{ij})}.$$

Le taux de coïncidence est égal à 1 si les configurations A et B sont confondues, et à une quantité petite, fonction de K, si elles sont disjointes. Ce critère est très aisé à programmer, et consomme beaucoup moins de temps que le calcul d'une distance euclidienne, par exemple. La constante K est nécessaire si l'on admet que deux ensembles vides sont identiques, et sa valeur est ajustée en fonction du seuil au-dessous duquel on considère une configuration comme nulle, ce qui permet de tenir compte du bruit de fond.

3.3. La segmentation de la parole en phonatomes et leur normalisation en temps

Appelons taux de stabilité $S(t)$ du squelette informatif à l'instant t le taux de coïncidence entre les deux spectres ou « événements » $E\left(t - \frac{\theta}{2}\right)$ et $E\left(t + \frac{\theta}{2}\right)$ séparés par une durée θ (distance d'ordre). La courbe de stabilité $S(t, \theta)$ est une caractéristique fondamentale du message; ses maximums indiquent les frontières des phonatomes (états quasi-stables), et ses minimums marquent les instants de plus grande instabilité, donc les événements les plus informatifs, que nous appelons événements-pivots. La figure 7 montre la courbe de stabilité obtenue sur un échantillon de parole (« Les petits »), et la segmentation-normalisation des phonatomes constituants : les phonatomes normalisés sont reconstitués par prélèvement de deux frontières, d'un pivot, et de deux événements intermédiaires. La segmentation observée sur la partie quasi-stable du /t/ n'est une erreur que si l'on se place au point de vue de la phonétique classique. En réalité ce locuteur a particulièrement accentué l'explosion du /t/, dont la partie quasi-stable est très voisine d'un /s/. Ceci peut être facilement corrigé à un stade ultérieur de la reconnaissance.

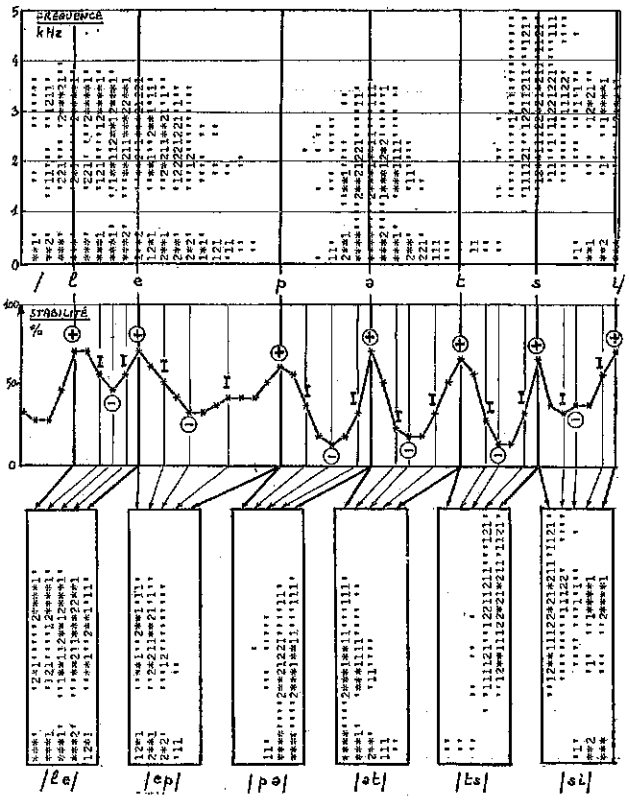


Fig. 7 — Principe de la segmentation du message en phonatomes, et de leur normalisation temporelle. Séquence « Les petits », locuteur masculin

Les seules fautes graves que l'on puisse observer au niveau de la segmentation sont constituées par la non détection d'une frontière. Le programme, testé sur un ensemble de phrases comprenant 145 phonatomes (nombre estimé lors de l'examen visuel des sonagrammes), n'a fourni que 3 erreurs de ce type, soit environ 2 %.

3.4. L'anamorphose fréquentielle

L'étude des sonagrammes de locuteurs divers, ainsi que certaines expérimentations avec la voix synthétique de l'Icophone [12, 13, 14, 15] nous ont permis de montrer que l'intelligibilité de la parole n'était que très peu affectée par un changement de l'échelle fréquentielle du squelette informatif. Entre une voix de basse et une voix d'enfant on peut constater un coefficient d'affinité en fréquence allant jusqu'à 2, ceci indépendamment de la hauteur de la « voix », c'est-à-dire de la fréquence vibratoire des cordes vocales.

Naturellement cette affinité n'est régulière qu'en première approximation. Mais compte tenu de l'imprécision attachée à la détermination des fréquences formantiques, un simple changement d'échelle permet d'augmenter notablement la ressemblance entre deux séquences de même contenu phonétique, issues de locuteurs différents morphologiquement. Le seul problème est de trouver le coefficient optimal d'anamorphose entre ces deux séquences.

Pour cela il nous faut comparer deux séquences normalisées en temps, afin de ne pas cumuler les difficultés. Nous n'avons conservé que les événements-frontières de la phrase « Marie ramasse une cerise » prononcée par un locuteur masculin et un locuteur féminin pris comme référence (fig. 8). Puis un coefficient d'anamorphose variant entre 0,5 et 1,5 par bonds de 0,05 a été appliqué à la séquence étudiée, et le taux de coïncidence a été calculé à chaque fois. Il apparaît ainsi que le taux de coïncidence passe par un maximum très net pour la valeur optimale du coefficient d'anamorphose.

La comparaison, répétée pour les 8 locuteurs pris deux à deux, a donné des résultats cohérents, qui rendent indiscutable la notion d'anamorphose fréquentielle. Cette notion, que nous étudions en synthèse depuis plusieurs années, nous a conduit à mettre

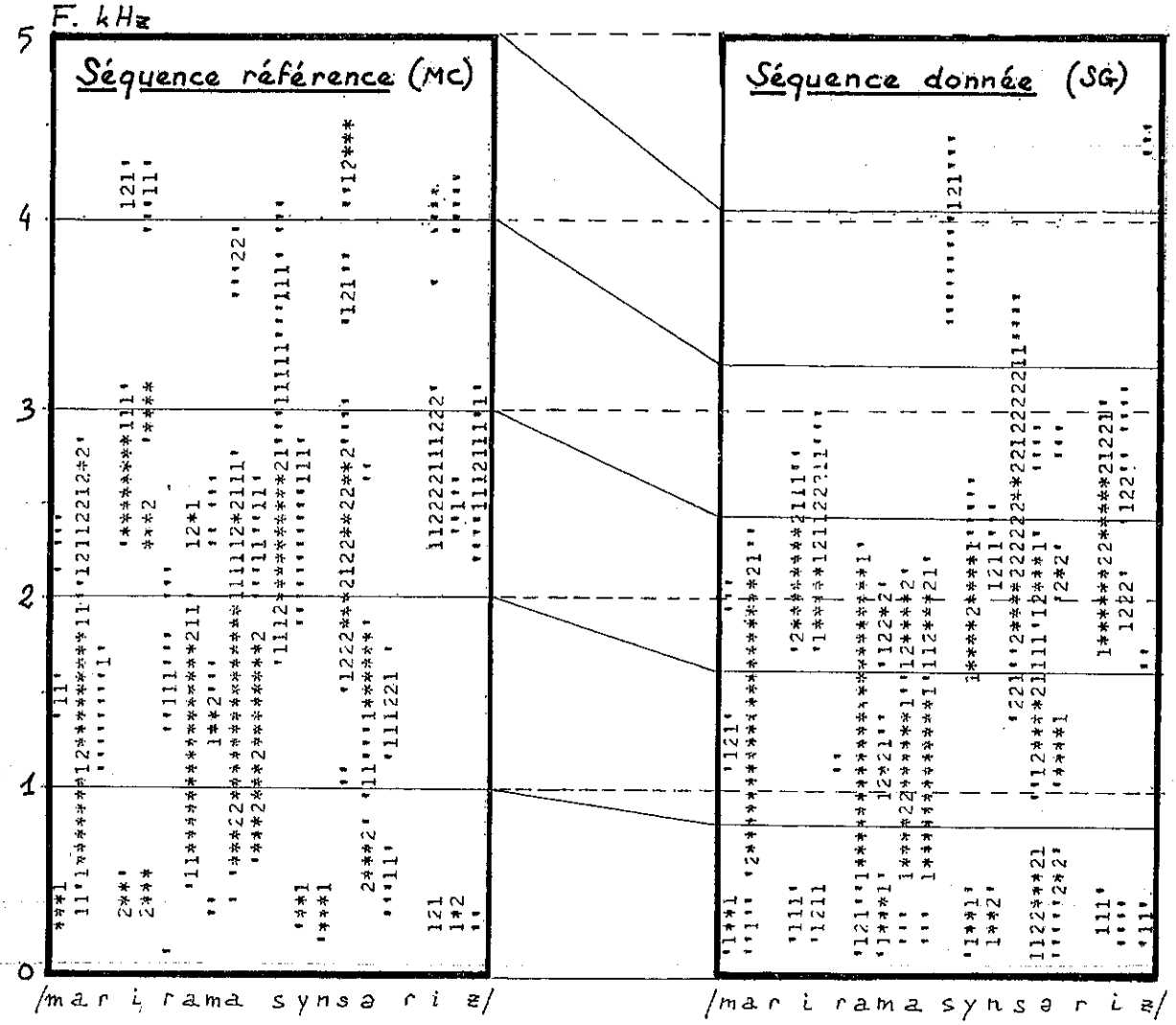
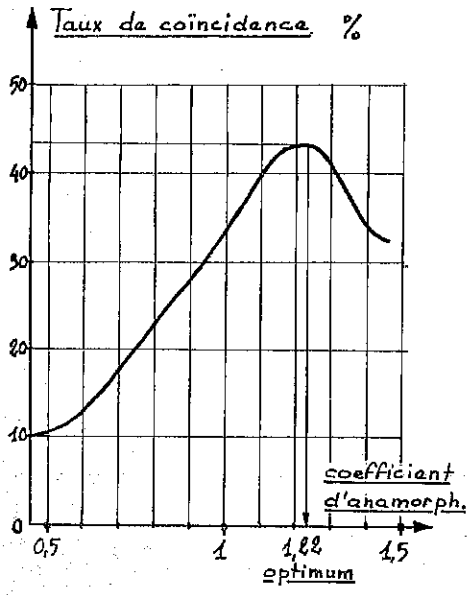


Fig. 8 — Recherche du coefficient optimum d'anamorphose régulière pour la voix d'un locuteur masculin (SG) par rapport à celle d'un locuteur féminin (MC)

Les deux séquences (« Marie ramasse une cerise »), normalisées en temps manuellement, sont représentées avec la même échelle fréquentielle. La seconde est anamorphosée en fréquence selon un coefficient variant de 0,5 à 1,5. Le taux de coïncidence passe par un maximum très net, pour le coefficient d'anamorphose optimum de 1,22. Les correspondances fréquentielles ainsi définies sont portées sur les sonagrammes.



au point un analyseur de parole dans lequel les fréquences centrales des filtres peuvent être toutes multipliées électroniquement par le même coefficient, de façon à effectuer la normalisation fréquentielle au niveau de l'analyse, et à décharger d'autant le calculateur recevant les données d'analyse (*).

(*) Etude poursuivie avec l'aide de la Direction des Recherches et Moyens d'Essais, dans le cadre des contrats 69/195 et 71/120.

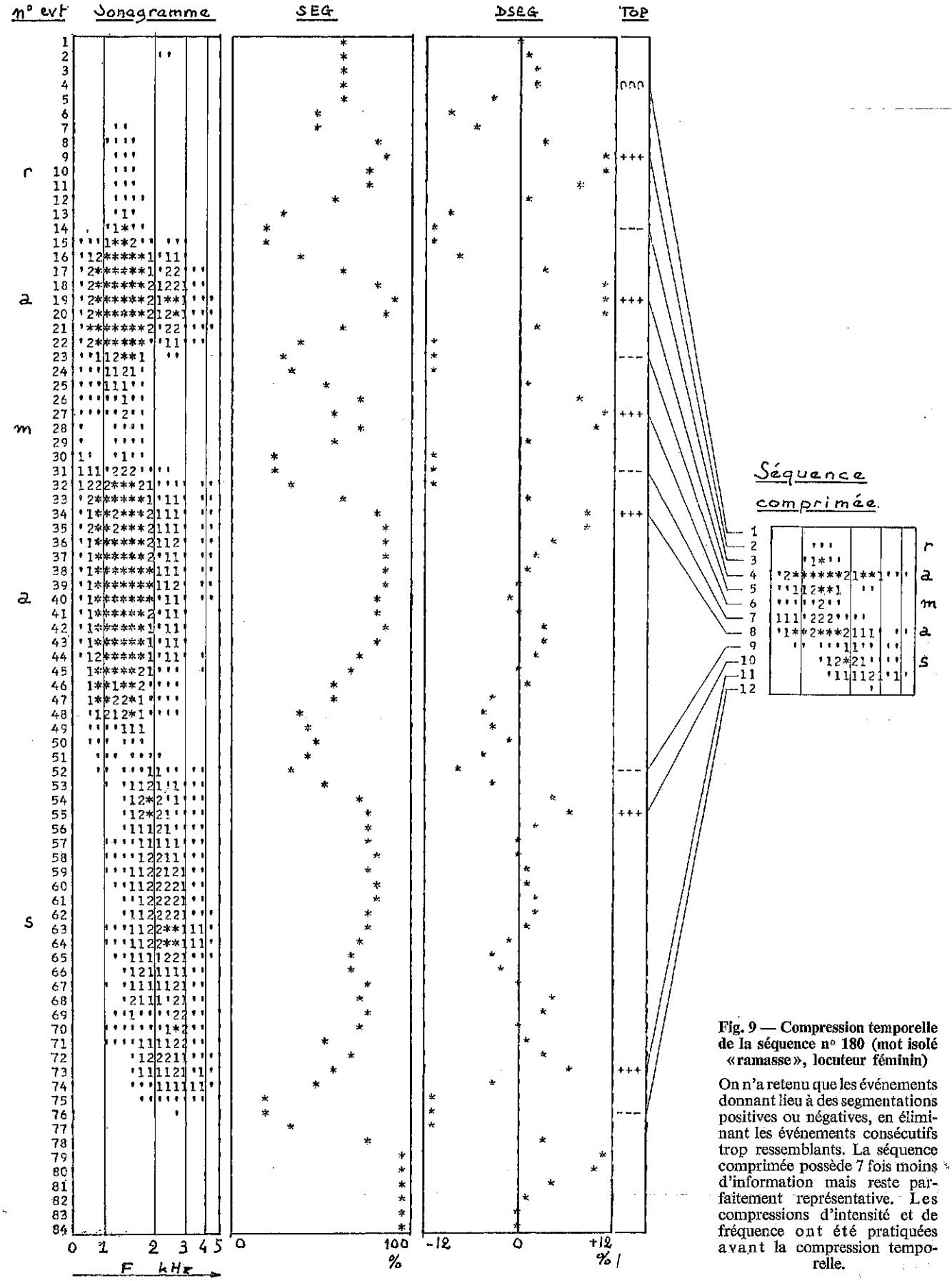


Fig. 9 — Compression temporelle de la séquence n° 180 (mot isolé «ramasse», locuteur féminin)

On n'a retenu que les événements donnant lieu à des segmentations positives ou négatives, en éliminant les événements consécutifs trop ressemblants. La séquence comprimée possède 7 fois moins d'information mais reste parfaitement représentative. Les compressions d'intensité et de fréquence ont été pratiquées avant la compression temporelle.

4. DEUX MÉTHODES DE RECONNAISSANCE

Nous décrirons maintenant deux expérimentations menées à partir des mêmes données acoustiques et illustrant, l'une la tendance analytique, l'autre la tendance globale.

4.1. Reconnaissance par phonatomes

Disposant d'un algorithme de normalisation des phonatomes, il était très simple d'en mémoriser un certain nombre selon un processus d'apprentissage et de les utiliser comme références dans une phase ultérieure de reconnaissance. Cette expérimentation a été menée pour un locuteur unique. Les 270 phonatomes composant les 7 phrases prononcées par ce locuteur ont permis de constituer un dictionnaire de 107 phonatomes différents, auxquels un nom a été affecté par l'opérateur. Ensuite les 108 phonatomes composant 13 mots prononcés quelques instants plus tard (nous avons pris grand soin de séparer les données d'apprentissage des données de reconnaissance) ont été identifiées avec un taux de réussite de 68 %. Le critère utilisé était ici encore le taux de coïncidence décrit plus haut.

Ce résultat, que nous ne donnons qu'à titre indicatif, est intéressant si l'on veut bien considérer deux points importants :

— Bien qu'issues d'un même locuteur, les séquences étudiées n'étaient soumises à aucune contrainte de prononciation, ne faisaient pas partie du même contexte phonétique et surtout n'étaient pas dirigées vers la reconnaissance.

— La suite de phonatomes n'est pas nécessairement le résultat intéressant en reconnaissance : il faut souvent la transformer en suite de phonèmes. Soit par exemple à reconnaître le mot « mirliton », composé des phonatomes « mi », « ir », « rl », « li », « it », « ton ». Un système qui reconnaîtrait parfaitement tous ces phonatomes donnerait en réalité une information redondante puisque la connaissance d'un phonatome sur deux (mis à part le début ou la fin) suffirait pour reconstituer le mot. On peut tirer parti de cette redondance lors du passage des phonatomes aux phonèmes, et augmenter ainsi de manière notable le taux de réussite.

4.2. Reconnaissance globale

La reconnaissance par mots semble devoir aboutir dans un avenir proche à des réalisations opérationnelles. Ses limites tiennent surtout aux volumes de calcul et de mémoire, croissant directement avec la taille du vocabulaire utilisable. Il est indispensable de réduire au minimum l'information superflue contenue dans chaque mot. Nous avons donc dans un premier temps étudié la compression d'information dans les trois dimensions d'intensité, de fréquence et de temps.

En intensité, nous avons limité à 8 (3 bits) le nombre de valeurs que peut prendre l'amplitude spectrale. En fréquence, les 64 voies d'analyse ont été ramenées à 16, réparties de manière non linéaire. La résolution la plus grande est placée aux alentours de 2 000 Hz.

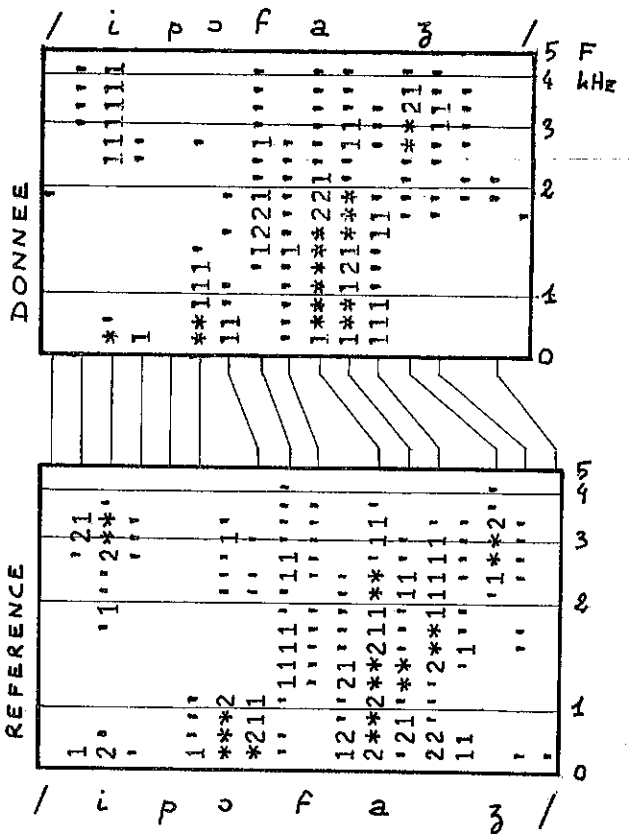


Fig. 10 — Reconnaissance globale par la méthode des correspondances (programmation dynamique)

La séquence n° 216 (mot « hippophage », locuteur féminin) est reconnue à partir de la séquence n° 76 (même mot, locuteur masculin). Les séquences sont comprimées en temps, fréquence et intensité. Il existe entre les deux séquences une anamorphose temporelle telle que leur simple juxtaposition donnerait un taux de coïncidence très faible. Une correspondance entre deux événements est établie lorsque leur taux de coïncidence est $\geqslant 40\,\%$; le taux moyen pour l'ensemble de la séquence passe ainsi à 50 %, et permet sa reconnaissance parmi les 14 mots-références.

En temps, enfin, la courbe de stabilité a été utilisée à nouveau, et les seuls événements retenus sont ceux qui correspondent à des extremums (fig. 9). De cette manière le débit d'information issu de l'analyse est divisé par 70 en moyenne, et n'est plus que d'un millier de bits par seconde à l'entrée du système de reconnaissance. On pourrait d'ailleurs augmenter encore ce rapport de compression, mais nous n'avons pas cherché à optimaliser le processus.

Nous avons appelé « méthodes de correspondances » la méthode de reconnaissance utilisée, qui s'apparente aux méthodes de programmation dynamique. Soit à comparer deux séquences X et R de même contenu phonétique, mais présentant entre elles des différences temporelles localisées ou réparties (fig. 10). Le programme compare initialement les événements X_1 et R_1. Si le taux de coïncidence C est inférieur à un seuil C_0, X_1 est comparé à R_2, puis R_3, etc... jusqu'à l'événement R_m pour lequel $C \geqslant C_0$. L'événement X_2 est ensuite comparé à R_{m+1}, puis si nécessaire à R_{m+2} etc... Si, au-delà de l'événement R_{m+M} on n'a pas trouvé d'événement R_k donnant un taux de coïncidence suffisant, on passe à l'événement X_3, et ainsi de suite.

On ne recherche donc les correspondances que dans une fenêtre temporelle de largeur M événements. Pratiquement nous avons fixé M à 5 événements; considérée en temps réel (avant compression), la largeur de la fenêtre est de l'ordre de 300 ms.

Le taux global de coïncidence est égal à la somme des taux partiels, rapportée à la longueur de la plus longue des deux séquences. Cette opération est effectuée pour chacun des mots-références; il suffit, pour achever la reconnaissance, de déterminer la référence pour laquelle on a obtenu le taux maximal.

L'ensemble de la reconnaissance ainsi conçue peut se dérouler très rapidement, car il n'est pas nécessaire d'attendre que l'acquisition de la séquence X soit terminée pour commencer le traitement.

Le programme a été testé sur 7 × 14 mots prononcés par 7 locuteurs masculins et féminins, les références étant les mêmes 14 mots prononcés une seule fois par le 8e locuteur. Le taux de reconnaissance obtenu s'est élevé à 70 %, chiffre extrêmement encourageant, compte tenu du fait que nous nous étions placés dans un cas de reconnaissance avec plusieurs locuteurs, sans adaptation ni apprentissage; ces conditions étaient beaucoup plus sévères que celles des expérimentations usuelles en reconnaissance vocale, qui portent généralement sur le même locuteur, avec un apprentissage plus ou moins prolongé.

5. CONCLUSION

La reconnaissance automatique de la parole est un sujet difficile, et passionnant dans la mesure où il n'est pas encore résolu. L'expérience acquise en analyse et en synthèse nous a naturellement conduit à explorer systématiquement un certain nombre d'options et d'idées nouvelles. Nous avons laissé de côté les méthodes mathématiques de classification, largement étudiées par ailleurs, pour nous intéresser exclusivement aux aspects acoustiques et perceptifs du phénomène de parole, qui nous semblent très loin en amont. Dans l'ensemble de l'étude nous n'avons aucunement cherché à réaliser des performances, qui nous auraient semblé prématurées, mais à dégager des principes. Maintenant commence un travail plus appliqué, et, peut-être, plus difficile.

BIBLIOGRAPHIE

[1] J.A. Dreyfus-Graf. — « Phonétographe et cybernétique ». Comptes rendus du 5e Congrès International d'Acoustique, Liège 1965.

[2] P. Alinat. — « Reconnaissance de phonèmes au moyen d'une cochlée artificielle ». *Automatisme*, tome XVII, n° 3, mars 1972.

[3] Baudry, Dupeyrat, Franck. — « Reconnaissance automatique de la parole ». — Colloque sur la parole organisée par le G.A.L.F. et l'A.F.C.E.T. au C.N.E.T., Lannion, mai 1972 (comptes rendus à paraître).

[4] J.Y. Gresser, G. Mercier. — « Exemple de reconnaissance automatique de la parole ». *Commutation et Electronique*, n° 32, janvier 1971.

[5] J. Caelen, S. Castan, G. Perennou. — « Reconnaissance de la parole en temps réel ». *Automatisme*, tome XVII, n° 3, mars 1972.

[6] C. Roche. — *Information utile en reconnaissance des formes et en compression de données. Application à la génération automatique de systèmes de reconnaissance optique et acoustique.* Thèse de doctorat d'Etat, Université Paris VI, décembre 1972.

[7] J.P. Tubach. — *Reconnaissance automatique de la parole.* Thèse de Doctorat d'Etat, Université de Grenoble, 1970.

[8] J.P. Haton, M. Lamotte. — « Prétraitement et reconnaissance de la parole. Simulation et réalisations pratiques ». *Automatisme*, tome XVII, n° 3, mars 1972.

[9] C. Gueguen, A. Maissis, L.F. Pau. — « Communication homme-machine sur support vocal ». *L'Echo des recherches*, octobre 1972.

[10] E. Leipp, M. Castellengo, J.-S. Lienard. — « La synthèse de la parole à partir de digrammes phonétiques ». Comptes rendus du 6e Congrès International d'Acoustique, Tokyo, août 1968.

[11] J.-S. Lienard, D. Teil. — « Les éléments phonétiques et la traduction automatique du message écrit en message parlé ». *Automatisme*, n° 10, octobre 1970.

[12] E. Leipp, J.-S. Lienard, M. Castellengo, J. Sapaly, D. Teil, A. Calinet, M. Mlouka. — « Colloque sur la parole ». Bulletin n° 53 du Groupe d'Acoustique Musicale de l'Université Paris VI, janvier 1971.

[13] E. Leipp, M. Castellengo, J. Sapaly, J.-S. Lienard. — « Les anamorphoses de la parole et leur simulation à l'ordinateur ». Comptes rendus du 7e Congrès International d'Acoustique, Budapest, août 1971.

[14] J.S. Lienard, M. Mlouka. — « Un aspect du prétraitement de la parole en vue de sa reconnaissance automatique ». Comptes rendus du 7e Congrès International d'Acoustique, Budapest, août 1971.

[15] J.-S. Lienard. — *Analyse, synthèse et reconnaissance automatique de la parole.* Thèse de doctorat d'Etat, Université Paris VI, avril 1972.

EXTRAIT DE LA REVUE « AUTOMATISME » N° 3, MARS 1973

DUNOD
ÉDITEUR

Imprimerie Bayeusaine, 14400 Bayeux