Zeitschrift: Bulletin technique de la Suisse romande
Band: 96 (1970)
Heft: 7: Foire de Bâle, 11-21 avril 1970

Artikel: Reconaissance automatique et synthèse de la parole
Autor: Benguerel, André-Pierre
DOI: https://doi.org/10.5169/seals-70847

Reconnaissance automatique et synthèse de la parole

par ANDRÉ-PIERRE BENGUEREL, ingénieur EPFL, Ph. D., professeur à l'Université de Colombie britannique

L'emploi de la parole pour la communication entre humains remonte à la préhistoire. Cependant, son étude en tant que phénomène physiologique, physique et perceptif n'a véritablement pris son essor que dans les années quarante. C'est aussi durant ces trente dernières années que des chercheurs, ingénieurs et linguistes, ont commencé à essayer de fabriquer des automates qui puissent « comprendre » la parole humaine, et « parler » et « être compris » par l'homme.

Il existe actuellement de nombreuses façons de communiquer avec une machine, mais toutes emploient une langue artificielle quelconque (p. ex. : un langage de programmation, dans le cas des ordinateurs). Il n'existe pas encore de machine, telle que HAL dans le film « 2001 », avec laquelle l'homme puisse communiquer dans sa langue, que ce soit oralement ou graphiquement (par l'emploi d'une écriture manuelle non stylisée). Cependant, les recherches dans ce domaine se sont considérablement intensifiées ces dernières années et le but de cet article [1] est de faire le point dans les deux domaines suivants :

1) la reconnaissance automatique de la parole ;
2) la synthèse de la parole.

Avant de pousser plus avant notre examen du problème de la reconnaissance et de la synthèse de la parole, il est indispensable de présenter tout d'abord quelques notions fondamentales de phonétique articulatoire et acoustique.

La chaîne parlée

La chaîne parlée peut être comparée, comme le montre la figure 1, à un schéma classique de télécommunication.

Le rectangle n° 1 de la chaîne correspond en première approximation aux domaines de la sémantique, de la syntaxe et de la neurophysiologie. En d'autres termes, c'est l'étude de la genèse du message, de sa structure, de sa relation avec les autres messages possibles, de son encodage en impulsions nerveuses et de leur transmission aux organes vocaux.

Le rectangle n° 2 correspond au domaine de la phonétique physiologique et articulatoire. C'est l'étude de la transformation *impulsions nerveuses – contractions musculaires*, de la production de la voix et des phénomènes qui l'accompagnent.

Le rectangle n° 3 correspond au domaine de la phonétique acoustique, c'est-à-dire à l'analyse de l'onde acoustique produite par l'appareil vocal.

Le rectangle n° 4 correspond au domaine de la phonétique perceptive. C'est l'étude de la réception de l'onde acoustique par l'oreille et de son décodage en impulsions nerveuses qui seront transmises au cortex cérébral.

Le rectangle n° 5 ressemble fortement au rectangle n° 1, à la différence près qu'ici, il ne s'agit pas de la formation et de l'encodage du message linguistique mais de son « déchiffrage ».

Phonétique

Dans la description des sons du langage, la phonétique articulatoire emploie plusieurs paramètres. L'un de ceux-ci est le *mode d'articulation*, c'est-à-dire la façon dont l'écoulement d'air est modifié par les organes vocaux, pour produire une certaine catégorie de sons.

La figure 2 montre une coupe schématique des organes vocaux.

Lorsque de l'air est expulsé des poumons, il s'échappe par la trachée, le pharynx, la bouche et éventuellement par le nez. Selon les obstacles qu'il rencontre, cet écoulement d'air est modulé de différentes façons. Si le passage pharyngo-oral ne comporte aucune constriction suffisante pour provoquer un changement de l'écoulement laminaire à l'écoulement turbulent, et si de plus la tension des cordes vocales est telle qu'elles sont en vibration, le son produit est une *voyelle*. Si le passage pharyngo-nasal est ouvert, la voyelle est dite *nasale*, s'il est fermé, la voyelle est dite *orale*.

Lorsque la constriction du passage pharyngo-oral devient plus importante, le son produit est une *consonne*. Si la constriction est telle que la turbulence commence à peine, la consonne est une *liquide*, une *semi-consonne* ou un *trille*. Si la constriction crée nettement une turbulence, la con-

[1] Le choix de ce bulletin pour une telle présentation pourra étonner certains lecteurs, mais il y a deux raisons principales à cela :

a) Une très petite partie seulement de la littérature pertinente paraît dans des revues d'ingénieurs. La majorité des publications se fait dans *The Journal of the Acoustical Society of America*, *Language and Speech*, *Phonetica*, *The Journal of Speech and Hearing Research*, etc.

b) La plus grande partie de cette recherche se fait aux Etats-Unis, en Suède, au Japon et en Grande-Bretagne. Il est donc improbable que beaucoup d'ingénieurs aient l'occasion de se familiariser avec cette recherche.

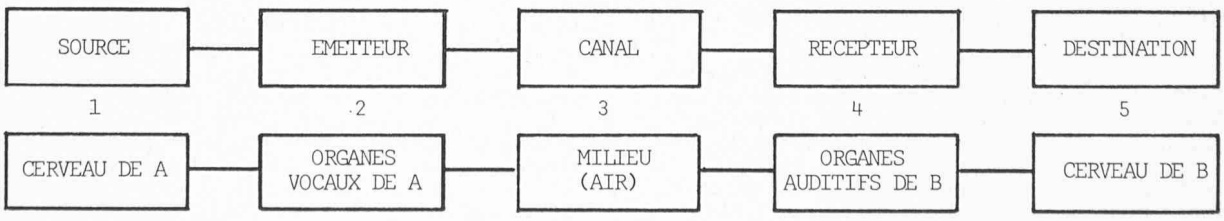

Fig. 1. — Analogie entre la chaîne parlée et un schéma de télécommunication.

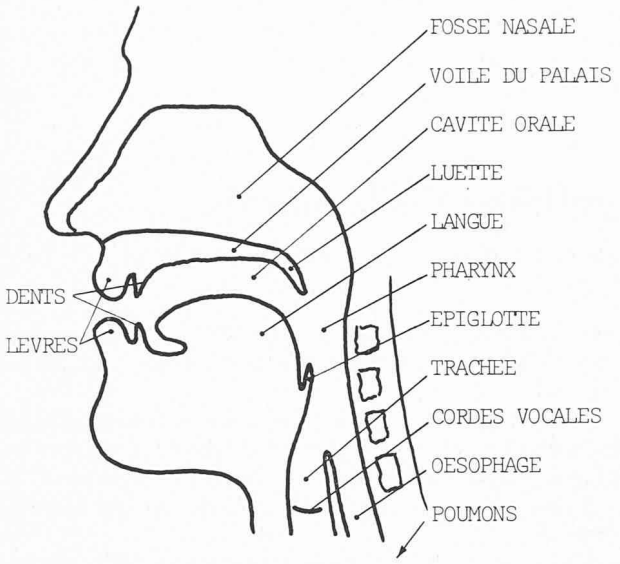

Fig. 2. — Coupe schématique des organes vocaux.

sonne est dite *fricative*. Si la constriction est telle qu'il y a occlusion orale complète, la consonne est dite *occlusive* (si le passage pharyngo-nasal est fermé), ou *nasale* (si le passage pharyngo-nasal est ouvert). Pour toutes les consonnes, il est possible d'avoir en plus la vibration des cordes vocales. Dans ce cas, la consonne est dite *voisée* ou *sonore* ; dans le cas contraire, elle est dite *sourde*.

Un autre paramètre nécessaire pour la description des sons, du point de vue articulatoire, est la position de la constriction, aussi appelée *lieu d'articulation*. Pour les voyelles, la grandeur de la constriction (si elle est assez faible pour ne pas produire de turbulence) ne détermine pas le son produit de façon univoque. La constriction doit être spécifiée par sa position horizontale (avant-centrale-arrière) et par sa position verticale (haute-milieu-basse). Le lieu d'articulation de diverses voyelles est schématisé dans la figure 3 [1].

Le lieu d'articulation des trois voyelles [i], [a], [u] est illustré par trois coupes sagittales de l'appareil vocal (fig. 4).

Le mode et le lieu d'articulation suffisent dans bien des cas à spécifier les voyelles d'une langue, mais dans le cas du français par exemple, deux paramètres supplémentaires sont nécessaires. Ils correspondent à la catégorie des voyelles *arrondies* (par opposition aux voyelles *écartées* ou non arrondies) : [y] (« p*u*r »), [ø] (« p*eu* »), [œ] (« p*eur* ») et à la catégorie des voyelles *nasales* (cf. plus haut) : [ɛ̃] (« v*in* »), [ã] (« v*ent* »), [õ] (« *on* ») et [œ̃] (« *un* »). Les voyelles [o] et [u], par exemple, sont également arrondies, mais il y a très peu de langues où il existe une voyelle arrière correspondante (en lieu et mode d'articulation) non arrondie.

Par contre, pour les voyelles avant, de nombreuses langues (le français, l'allemand, le suédois par exemple) ont un ensemble de voyelles non arrondies [i], [e], [ɛ] et un ensemble correspondant de voyelles arrondies, respectivement [y], [ø], [œ].

Le lieu d'articulation est également un paramètre important pour la description des consonnes. Cependant, pour les consonnes, la position verticale de la constriction est déjà implicite dans la spécification du mode d'articulation (e.g. occlusive, fricative, trille). Par conséquent, le lieu d'articulation des consonnes est spécifié par sa position avant-arrière seulement, mais avec un plus grand nombre de possibilités que dans le cas des voyelles. Par exemple [1] :

bilabiale :	[m] (« *m*a »), [p] (« *p*as »), [b] (« *b*u »), [w] (« *w*eek-end »), [ɥ] (« *hu*it »)
labiodentale :	[f] (« *f*ut »), [v] (« *v*a »)
linguadentale :	[θ] (anglais « *th*ick »), [ð] (anglais « *th*e »)
alvéolaire :	[n] (« *n*u »), [t] (« *t*u »), [d] (« *d*u »), [s] (« *s*u »), [z] (« *z*ut »), [l] (« *l*a »), [r] (italien « *R*oma »)
palatale :	[ʃ] (« *ch*ou »), [ʒ] (« *j*eu »), [j] (« *i*ode »)
vélaire :	[k] (« *qu*i »), [g] (« *g*are »)
uvulaire :	[R] (« t*r*ois »)

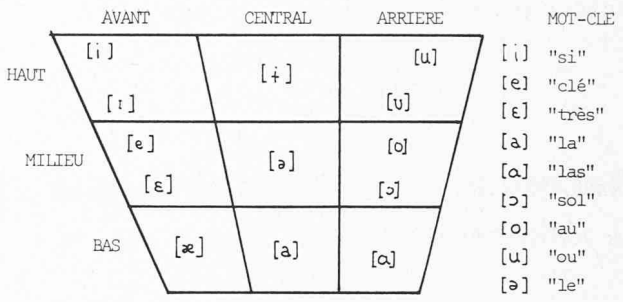

Fig. 3. — Représentation schématique du lieu d'articulation de quelques voyelles.

Phonémique

L'appareil vocal humain peut produire une infinité de sons, par le simple fait que certains paramètres articulatoires varient de façon continue plutôt que discrète (ex. : le lieu d'articulation). La plupart de ces différences ne dépassent pas le seuil de perception. De surcroît, plusieurs différences beaucoup plus importantes et nettement perceptibles ne sont pas exploitées systématiquement par les locuteurs d'une certaine langue. Par exemple, le contraste *voyelle orale – voyelle nasale* n'est utilisé que par un petit nombre de langues (dont le français, le portugais et le polonais).

La question se pose donc : quels sont les contrastes qui, pour une langue donnée, sont significatifs ? En d'autres termes, si l'on veut représenter de façon économique le langage parlé, est-il toujours nécesaire d'utiliser les symboles phonétiques ou existe-t-il une représentation plus économique, mais qui néanmoins apportera l'information nécessaire à l'auditeur ou au lecteur ? La réponse est affirmative : cette représentation peut se faire au moyen de phonèmes [2].

Un phonème est un ensemble (au sens mathématique) de sons qui ne contrastent pas linguistiquement entre eux,

[1] Par convention, les symboles phonétiques sont utilisés entre crochets [ ], afin de bien les différencier des symboles orthographiques ou lettres. En effet, la correspondance *son – symbole phonétique* est biunivoque, alors que la correspondance *lettre-symbole phonétique* (en français en tout cas) ne l'est pas. Par exemple, les mots « haut », « au », « aux », « eau », « oh », « os » seront tous transcrits par [o]. Un mot clé est donné, en forme orthographique, pour chaque son qui peut être utilisé en français.

[1] Cette liste, très incomplète, omet de nombreux sons utilisés dans des langues autres que le français et qu'il serait vain de vouloir énumérer ici sans illustration sonore.

[2] Les linguistes ne se sont pas encore mis complètement d'accord sur la définition du phonème, et en particulier sur son statut dans la hiérarchie des niveaux linguistiques. La position présentée par Postal et par Chomsky et Halle est de loin la plus solide ; toutefois, pour ne pas allonger inutilement cet article mais sans risquer d'être moins général, c'est la notion traditionnelle qui est présentée ici.

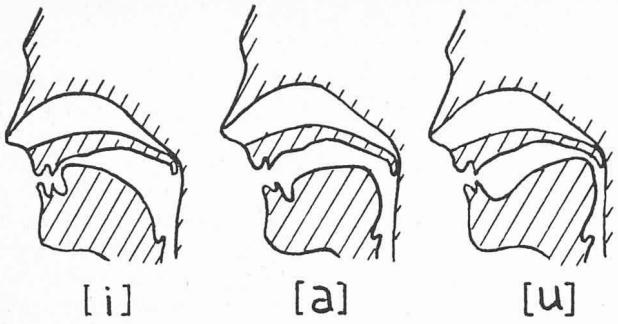

Fig. 4. — Coupes sagittales de l'appareil vocal pour les voyelles [i], [a] et [u].

mais qui contrastent avec n'importe quel autre son qui n'appartient pas à cet ensemble. Pour un locuteur arabe, par exemple, le son [c] (occlusive palatale sourde ; telle qu'au début de « qui » en français) et le son [k] (occlusive vélaire sourde, telle qu'au début de « cou » en français) contrastent dans tous les cas. Par conséquent on parlera du phonème arabe /c/ et du phonème arabe /k/ [1].

En français, par contre, bien que les sons initiaux de « *q*ui » et de « *c*ou » soient phonétiquement différents, le locuteur n'y prête guère attention, et pour lui ces deux sons ([c] et [k]) appartiennent au même phonème. En notation mathématique :

en arabe : /c/ = { [c] }
/k/ = { [k] }
en français : /k/ = { [c], [k] } [2]

En linguistique, on dit que le phonème français /k/ a deux allophones, [c] et [k], alors que les phonèmes arabes /c/ et /k/ n'ont qu'un allophone chacun, [c] et [k] respectivement.

Un allophone est un sous-ensemble de sons linguistiquement non différenciés. La différence entre deux allophones (p. ex. : [c] et [k] en français) est simplement plus grande (phonétiquement) que celle entre deux sons appartenant au même allophone (p. ex. : un [c] palatal et un [c] prépalatal), mais ni l'une ni l'autre ne constitue un contraste phonémique (dans la langue en question).

Il devrait être évident que la notion de phonème n'a de sens que dans une langue, ou même dans un dialecte particulier. La notation /k/, dans l'exemple ci-dessus, est ambiguë si l'on ne spécifie pas de quelle langue il s'agit.

Dans l'évolution d'une langue, il y a apparition et disparition de phonèmes. Actuellement, en français, il y a une tendance vers la fusion de /$\tilde{\varepsilon}$/ et de /$\tilde{\text{œ}}$/ en un seul phonème. Pour une grande partie des locuteurs du centre et du nord de la France, les mots « brin » et « brun » par exemple ne sont plus différenciés dans la prononciation. Dans le sud et l'est de la France et en Suisse romande, par contre, le contraste est encore valable pour la majorité des locuteurs. Le concept de phonème /$\tilde{\varepsilon}$/ ou de phonème /$\tilde{\text{œ}}$/ ne s'applique donc pas à tous les dialectes du français de façon identique. Remarquons aussi que (malheureusement) les ensembles de sons constituant les phonèmes ont rarement une intersection nulle. En d'autres termes, il arrive souvent qu'un son appartienne à deux ou trois phonèmes. Comme nous le verrons plus loin, c'est là une des sources majeures de difficultés dans le problème de la reconnaissance automatique de la parole.

[1] Par convention, les symboles phonémiques sont utilisés entre barres obliques, par exemple /p/, /k/, etc., afin de les différencier des symboles phonétiques.

[2] Puisqu'il y a forcément moins de symboles phonémiques que de symboles phonétiques, le choix du symbole /k/ pour représenter le phonème [c], [k] est évidemment arbitraire. N'importe quel autre symbole (/c/ par exemple) conviendrait tout aussi bien.

Phonétique acoustique

Au lieu de décrire la façon dont les sons du langage sont produits, il est également possible d'analyser l'onde acoustique produite par l'appareil vocal. Si l'on examine simplement l'onde produite au moyen d'un oscillographe, par exemple, il est très difficile d'en tirer des caractéristiques intéressantes.

Cependant, puisqu'il s'agit maintenant d'un problème d'acoustique, il est possible d'utiliser les instruments et les techniques de la physique. L'une de ces techniques est l'analyse spectrale.

L'examen de l'onde parlée dans le domaine de la fréquence a deux avantages : premièrement, l'analyse de l'appareil vocal du point de vue acoustique montre que le concept de fréquence propre permet une description concise des sons du langage ; deuxièmement, les études physiologiques de l'audition ont prouvé qu'au stade initial tout au moins, l'oreille opère une analyse fréquentielle grossière dans le traitement de l'onde acoustique. Si l'analyse fréquencielle se montre utile dans l'étude de la production et de la perception de la parole, il semble naturel de l'employer également au niveau acoustique, ce que l'expérience confirme.

Il existe trois types fondamentaux d'ondes parlées : les ondes quasi périodiques, les ondes aléatoires, et les impulsions.

Les ondes quasi périodiques ont un spectre constitué d'harmoniques dont la fondamentale correspond à la fréquence de vibration des cordes vocales. L'enveloppe de ce spectre présente des maxima que l'on appelle *formants*. Le premier formant correspond au maximum situé à la fréquence la plus basse, et ainsi de suite. Chaque formant correspond à une résonance de l'appareil vocal. Ainsi lorsque le son produit passe d'un [a] à un [u] sur le même ton, les formants changent de fréquence, alors que la séparation entre harmoniques reste identique (cf. fig. 5).

Au contraire, lorsque le son produit varie d'un [a] chanté sur un la_{110} à un [a] chanté sur un la_{220}, les formants ne changent pas de fréquence mais la distance entre harmoniques double (cf. fig. 6).

Les ondes du type aléatoire sont caractérisées par un spectre continu très semblable à un spectre de bruit. Les différences entre sons correspondent à des distributions différentes de l'énergie dans le spectre.

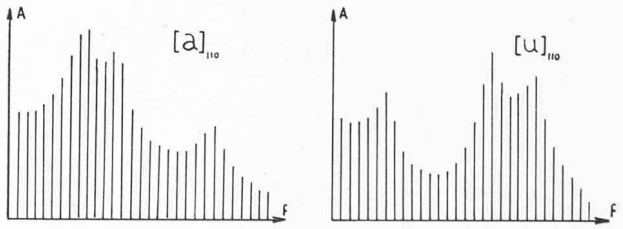

Fig. 5. — Spectre des voyelles [a] et [u] prononcées à une fréquence fondamentale de 110 Hz.

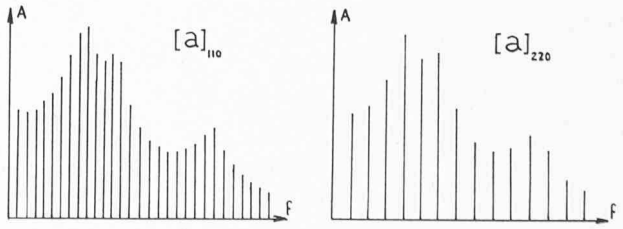

Fig. 6. — Spectre de la voyelle [a] produite à une fréquence fondamentale de : *a*) 110 Hz ; *b*) 220 Hz.

Le son [s], par exemple, présente une énergie distribuée sur une bande de fréquence beaucoup plus large que le son [ʃ]. Dans le cas des consonnes fricatives, les antirésonances (correspondant aux zéros de l'impédance acoustique de transfert de l'appareil vocal) sont souvent plus instructives quant au son produit que les résonances (qui correspondent aux pôles de l'impédance de transfert).

Les ondes du type impulsion sont également caractérisées par un spectre continu à bande plus ou moins large. La différence avec une onde du type aléatoire est dans la nature de l'onde plutôt que dans celle du spectre. L'impulsion a une durée assez limitée, et généralement elle est précédée d'un silence (cas des occlusives sourdes), ou d'une onde quasi périodique à intensité assez faible (cas des occlusives sonores). L'onde produite par le son [p] par exemple, comprend une première partie à amplitude quasi nulle (correspondant à l'occlusion de la bouche et au silence résultant) suivie d'une impulsion assez courte (de l'ordre de 20 à 30 ms) qui correspond au relâchement brusque de la pression orale qui avait augmenté pendant l'occlusion.

Tous les sons du langage ne sont pas associés à un type unique d'onde acoustique. De nombreux sons produisent une onde qui est une combinaison des types de base décrits plus haut. Le son [z] par exemple est une combinaison d'une onde quasi périodique et d'une onde aléatoire.

Les laboratoires de recherches de la Bell Telephone ont mis au point, pendant la deuxième guerre mondiale, un instrument appelé *spectrographe de sons*. Cet instrument inscrit, sur un papier spécial, un spectogramme, c'est-à-dire une analyse spectrale en fonction du temps de 2,4 secondes de parole. Le principe de fonctionnement de l'appareil est décrit très schématiquement dans la figure 7.

Un échantillon de parole de 2,4 secondes est tout d'abord enregistré sur un tambour magnétique. Cet échantillon est ensuite reproduit 200 fois au travers d'un filtre de bande. A chaque répétition de l'échantillon, la fréquence du centre de la bande passante du filtre est modifiée de façon telle qu'à la fin de la 200e répétition, la bande de fréquence du signal original ait été complètement parcourue.

(Le résultat est équivalent à l'analyse simultanée par une série de filtres de bande en parallèle). En fait, il est plus pratique d'utiliser un filtre à bande passante fixe et de faire « glisser » le spectre du signal à analyser « devant » le filtre. Pour cela, on module le signal avec une porteuse à haute fréquence et on « glisse » une bande latérale du signal « devant » le filtre fixe. La translation s'opère par variation de la fréquence porteuse. Le contrôle de cette dernière est couplé mécaniquement au tambour sur lequel est enregistré le signal. Il est aussi couplé mécaniquement au déplacement du dispositif d'écriture (marqueur). Celui-ci reçoit un courant proportionnel à l'intensité du signal sortant du filtre analyseur, et brûle un papier conducteur proportionnellement au courant. Grâce au couplage mécanique, porteuse et marqueur parcourent simultanément la bande de fréquence du signal, produisant ainsi un spectrogramme. Deux largeurs de bande sont employées : 300 Hz ou 45 Hz. Le spectrogramme à bande étroite (45 Hz) permet une meilleure résolution en fréquence : les harmoniques sont clairement visibles. La résolution dans le domaine du temps, par contre, n'est pas très bonne. Le spectrogramme à bande large (300 Hz), au contraire, permet une bonne résolution dans le temps : on peut distinguer chaque impulsion glottale sous forme de stries verticales. Dans le domaine de la fréquence, la résolution n'est pas suffisante pour distinguer les harmoniques ; cependant les formants sont en général plus facilement discernables que sur un spectrogramme à bande étroite. Le choix du filtre dépend donc en grande partie de genre d'information recherché. Les figures 8 et 9 représentent chacune un spectrogramme, à bande étroite et à bande large respectivement. Dans la figure 8, les changements de fondamentale sont bien visibles (ils correspondent en première approximation aux changements d'intonation). Dans la figure 9, les changements de fréquence des formants sont très distincts, mais les variations de fondamentale ne sont plus directement visibles.

Indirectement toutefois, on peut remarquer que les stries sont plus serrées (vu la période de vibration glottale plus courte) lorsque la fondamentale est élevée.

On peut aussi remarquer que lorsque la position d'articulation d'une voyelle se déplace d'avant en arrière (de [i] à [a] par exemple), la fréquence du deuxième formant (F_2) diminue. Pareillement, lorsque la position d'articulation de la voyelle passe de haut en bas (de [i] à [a] par exemple), la fréquence du premier formant (F_1) augmente. En fait,

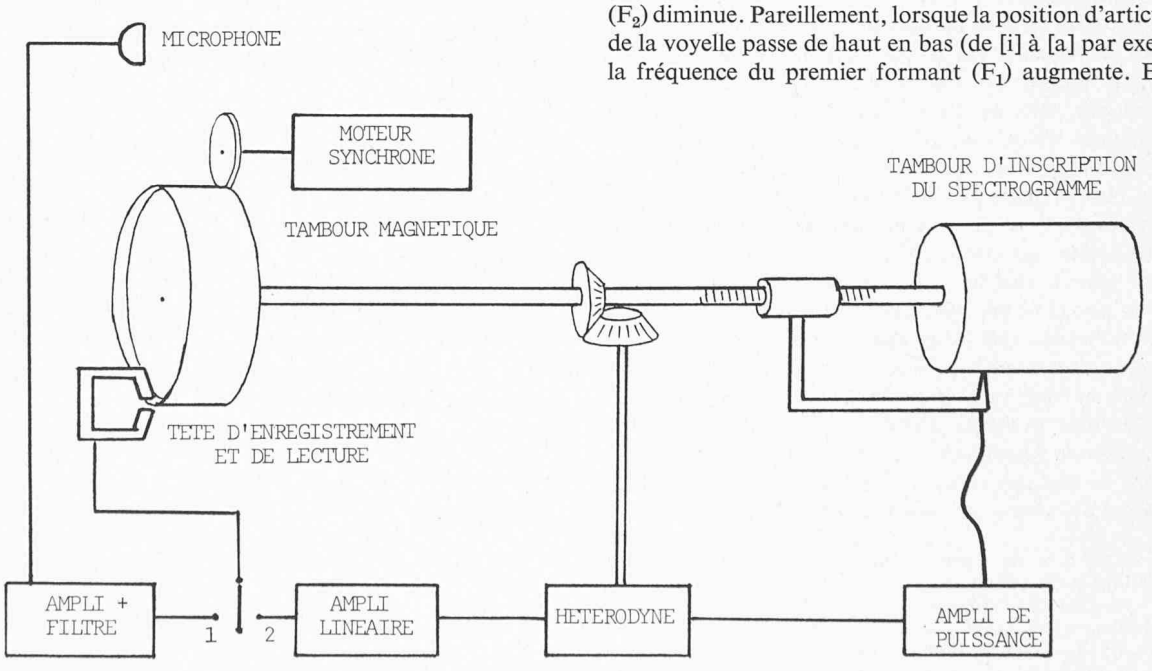

Fig. 7. — Schéma de principe du spectographe BTL.

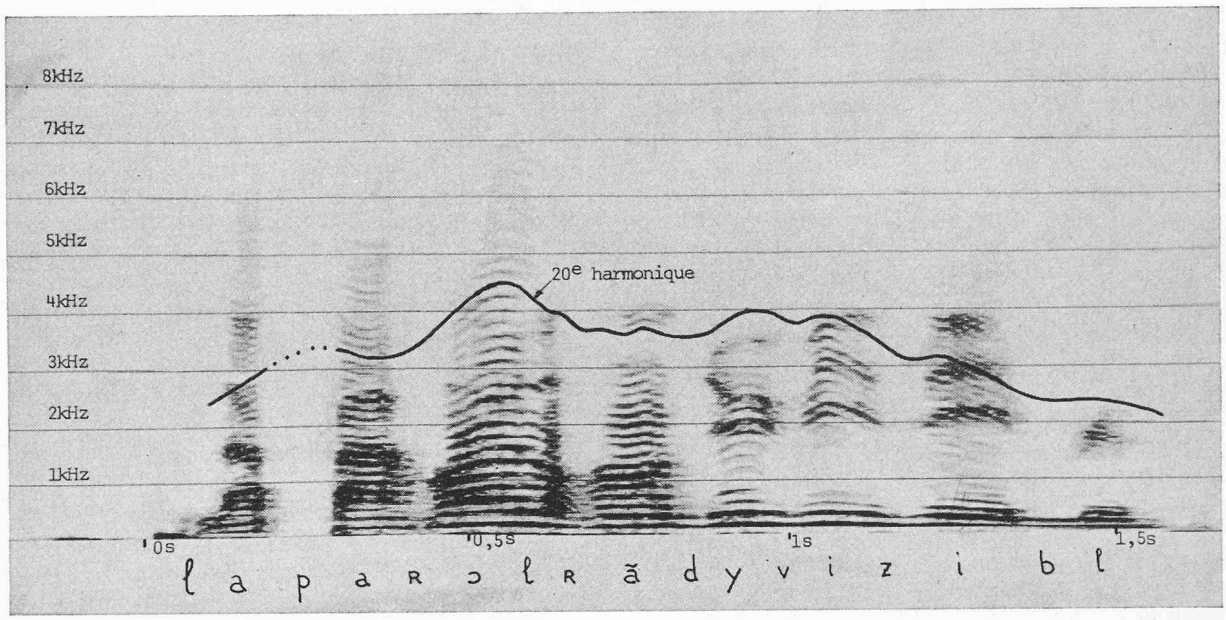

Fig. 8. — Spectrogramme à bande étroite de la phrase « la parole rendue visible ».

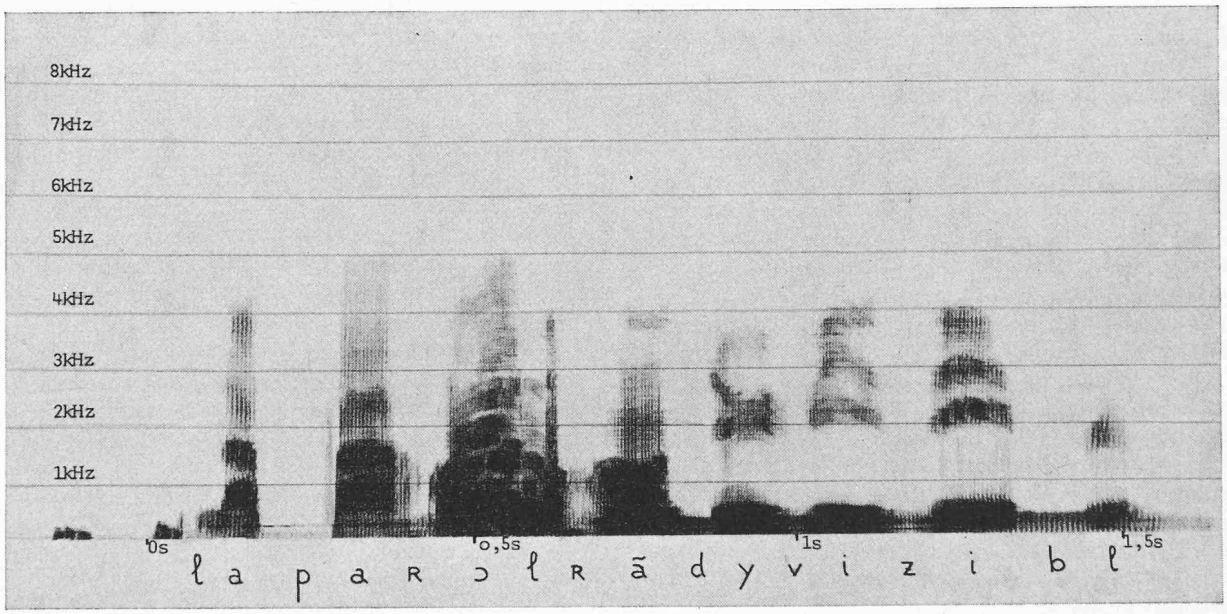

Fig. 9. — Spectrogramme à bande large de la même phrase.

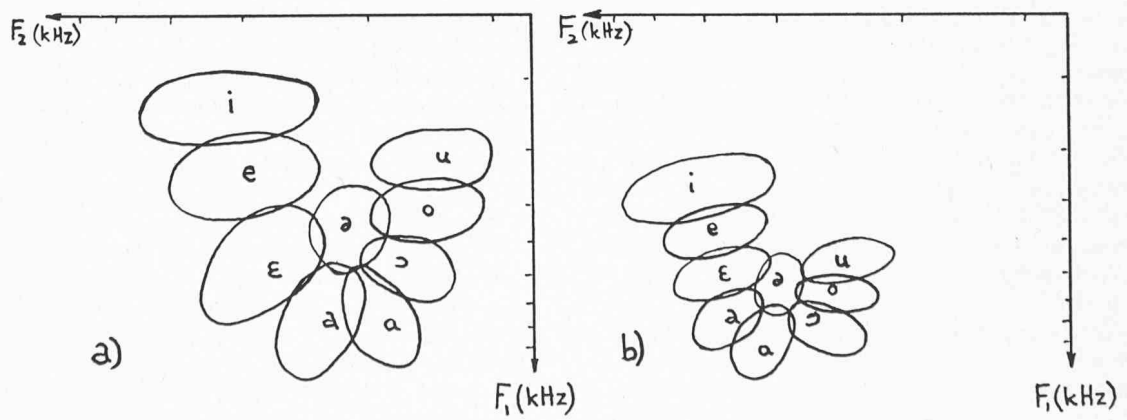

Fig. 10. — Diagrammes F_1/F_2 : *a*) pour les voyelles d'un locuteur adulte ; *b*) pour les voyelles d'un enfant.

si l'on mesure F_1 et F_2 pour plusieurs séries de voyelles soutenues et qu'on reporte le point (F_1, F_2) pour chaque voyelle, on obtient le diagramme de la figure 10 a). En comparant la figure 10 a) et la figure 3, on remarque immédiatement la similitude dans la disposition des voyelles dans les deux diagrammes. Cette ressemblance n'est pas une pure coïncidence ; cependant, en vingt ans de recherches, il n'a pas encore été possible d'éclaircir entièrement la relation entre ces deux diagrammes. Si l'on trace un diagramme F_1/F_2 tel que celui de la figure 10 a), mais pour un enfant, le diagramme des voyelles subit une translation vers des valeurs de F_1 et F_2 plus grandes. Ceci s'explique par le fait qu'un enfant, ayant des dimensions physiques plus petites, les fréquences propres de son appareil vocal sont par conséquent plus élevées. Le [u] d'un enfant (dans un diagramme F_1/F_2) correspond donc approximativement au [ɛ] d'un adulte. Ceci suggère donc immédiatement que la reconnaissance des voyelles, si elle est basée sur l'extraction des formants de l'onde acoustique, se fait sur un schéma nettement plus complexe que celui suggéré par la figure 10. Nous allons examiner maintenant le problème de la reconnaissance de la parole et de ses insuccès (ou succès selon le point de vue) jusqu'à aujourd'hui.

Reconnaissance automatique de la parole

La construction d'une machine qui puisse reconnaître la parole constitue un problème des plus complexes, impossible serait-on tenté de dire, s'il n'existait pas déjà une machine qui en soit capable : l'homme. Dans le cas idéal, une telle machine reçoit un message linguistique sous forme d'une onde acoustique et le transforme en une séquence de phonèmes. Un locuteur francophone peut sans grande difficulté écouter un message en français et le transcrire phonémiquement. A partir de la transcription phonémique, il est toujours possible de passer au stade orthographique. La reconnaissance des éléments linguistiques est basée sur la connaissance des contraintes contextuelles phonologiques, syntaxiques et sémantiques de la langue en question. La reconnaissance automatique de la parole sous-entend l'analyse phonémique, et l'analyse phonémique sous-entend la connaissance (et l'emploi) de ces contraintes. Il est certainement possible de simuler grossièrement le genre d'analyse effectué par l'oreille humaine, mais à ce jour aucune des machines construites, même la plus complexe, n'a démontré une aptitude à utiliser les contraintes linguistiques comparable, même de loin, à celle de l'homme. On ne saurait trop insister sur la différence, le gouffre pourrait-on dire, qui existe entre une reconnaissance phonétique et une reconnaissance phonémique de la parole. Dans le premier cas, la seule condition est que le message soit produit par l'appareil vocal humain. Dans le second cas, on présuppose une connaissance complète de la langue (phonologie, syntaxe, sémantique). La reconnaissance phonétique de la parole est dans le domaine des possibilités pour les techniques actuelles de l'analyse de la parole, mais la reconnaissance phonémique en est encore très loin.

Si la reconnaissance (phonémique) de la parole devient un jour possible à un prix abordable, les applications ne manqueront pas. Les numéros de téléphone n'auront plus besoin d'être composés au moyen d'un disque ou de boutons-poussoirs. Il suffira d'énoncer le numéro voulu, ou peut-être même simplement de dire le nom et l'adresse de l'abonné désiré. De même, l'accès aux ordinateurs et leur programmation pourront se faire directement par la voix, sans avoir recours aux cartes ou aux rubans perforés ou encore aux téléscripteurs.

Revenons sur terre et examinons quelques-unes des tentatives de reconnaissance de la parole.

Le système imaginé par Davis, Biddulph et Balashek (1952) et baptisé « Audrey » utilise un principe assez simple pour reconnaître les chiffres 0 à 9. Les formants F_1 et F_2 sont évalués en fonction du temps. La trajectoire temporelle du point (F_1 ; F_2) dans le plan F_1F_2 est alors comparée, après normalisation en amplitude et en durée, aux dix trajectoires de référence stockées dans la mémoire de l'ordinateur. La trajectoire de référence offrant la meilleure corrélation avec la trajectoire inconnue est alors choisie comme étant la réponse correcte. La machine ne comporte aucun ajustement automatique à la voix du locuteur. Cet ajustement doit se faire manuellement. Ceci fait, le pourcentage de réponses correctes pour un signal (l'un des dix chiffres) ayant la fidélité d'une conversation téléphonique varie entre 97 % et 99 %.

Une technique semblable a été utilisée par Dudley et Balashek (1958), pour essayer de reconnaître dix sons (six voyelles et n, r, f, s). Au lieu d'étudier la trajectoire du point (F_1 ; F_2), ils examinent les sorties simultanées de dix filtres de bande (300 Hz) contigus et recouvrant la largeur de bande du signal (3000 Hz). La corrélation entre les spectres des sons de référence et le son à identifier est calculée de façon continue, et la paire *référence-inconnue* offrant la corrélation maximum donne une indication sur le son à analyser. Pour ensuite reconnaître un chiffre à partir des sons qui le composent, ce système normalise, tout comme Audrey, l'amplitude et la durée du signal donné. Par un système de circuits RC couplés, dix condensateurs accumulent chacun une charge proportionnelle à la corrélation entre l'un des dix signaux de référence et le signal inconnu. A la fin du signal, le condensateur ayant la charge la plus grande est choisi comme la réponse correcte la plus probable. Ce système, plus perfectionné que le précédent, donne également de bons résultats lorsqu'il est ajusté pour une voix particulière.

Un autre système est celui de Fry et Denes, qui utilise également des spectres de référence. La comparaison toutefois se fait de façon différente. La sélection d'un phonème pour un certain segment est déterminée par la vitesse de variation de la configuration spectrale.

De plus, pour la première fois, certaines contraintes linguistiques sont utilisées pour l'identification : il est tenu compte des probabilités conditionnelles pour qu'un phonème β suive un phonème α (les cas limites étant l'impossibilité ou au contraire la certitude que β suive α). L'inconvénient de cette technique est qu'une fois embarqué sur une fausse piste, à cause d'une erreur initiale, il y a peu d'espoir de retour. Dans le système de Fry et Denes (fig. 11), de sévères restrictions sur la grandeur du vocabulaire utilisé préviennent en grande partie cette éventualité. 14 phonèmes peuvent être reconnus.

Lorsque les contraintes linguistiques mentionnées ne sont pas utilisées, le pourcentage d'identifications correctes pour des sons et pour des mots est de 60 % et 24 % respectivement. Lorsque les contraintes linguistiques sont utilisées, ces pourcentages montent à 72 % et 44 % respectivement. Lorsque un second et un troisième locuteurs sont utilisés, sans réajuster le système, le pourcentage (pour les sons) retombe à 45 % environ.

Si l'on veut inclure des données linguistiques dans le système de reconnaissance, le stockage et le traitement de cette information deviennent immédiatement très complexes. Le calculateur digital devient l'outil rêvé pour ce genre de travail ; en fait, une fois le signal mis sous forme digitale, la suite de la reconnaissance peut être entièrement

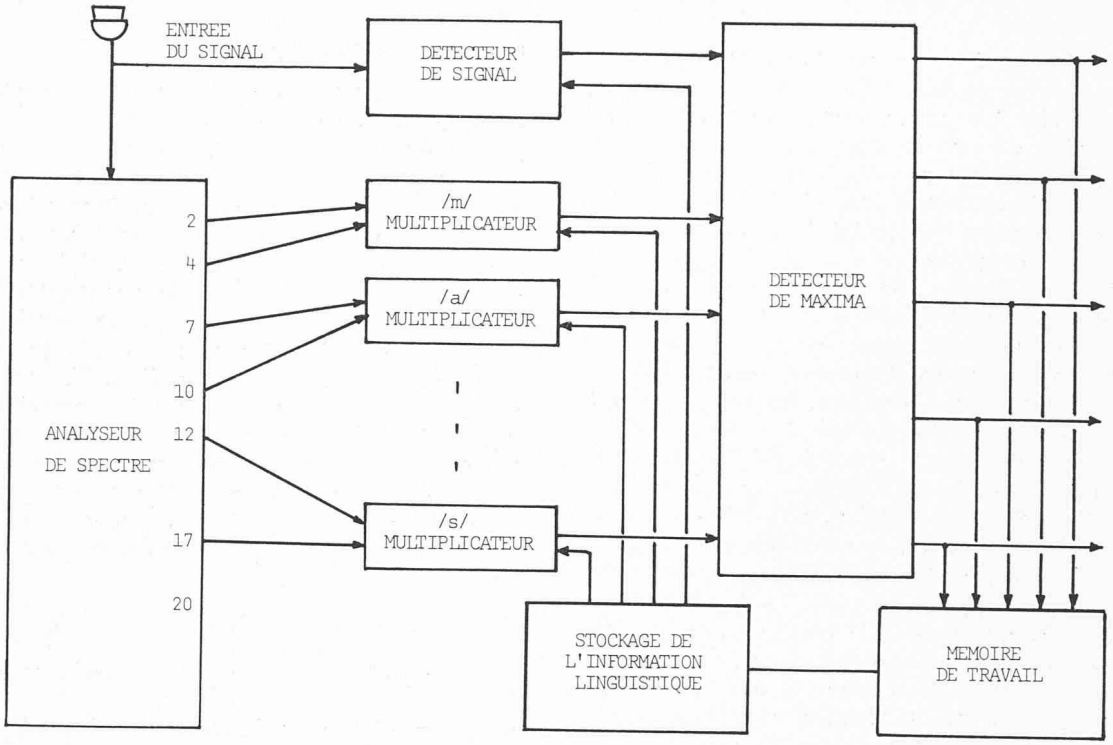

Fig. 11. — Schéma de principe du système de Fry et Denes.

effectuée par l'ordinateur, y compris le filtrage (digital) pour obtenir le spectre.

Nous ne mentionnerons pas les nombreuses tentatives, souvent très semblables, d'autres chercheurs. Pratiquement, leur point de départ est toujours identique : l'analyse spectrale du signal, qui correspond très approximativement à la reconnaissance phonétique du message. A partir de là, les stratégies diffèrent, mais malheureusement se ressemblent en ce qui concerne le résultat final : aucun système ne fonctionne pour un grand nombre de voix, ni pour un grand vocabulaire, alors que l'homme, lui, réussit presque parfaitement dans l'une et l'autre tâches.

Des sommes considérables sont consacrées à la recherche dans ce domaine, en grande partie, semble-t-il, à cause des applications séduisantes qui découleraient de la solution du problème. Il devient cependant de plus en plus évident que la recherche doit d'abord élucider certaines questions fondamentales, telles que la perception et l'analyse de la parole par l'homme, avant de s'attaquer à ce qui est plus ou moins l'étape finale du problème. La reconnaissance automatique de la parole, telle que l'homme la réalise, ne sera possible que par l'analyse adéquate de la structure du langage, non seulement des points de vue phonologique, syntaxique et sémantique, mais également des points de vue physiologique et neurophysiologique.

Un des outils les plus utiles dans l'étude systématique de la perception et de la reconnaissance de la parole par l'homme est le synthétiseur de parole. C'est donc vers ce sujet que nous allons nous tourner maintenant.

Synthèse de la parole

Alors que pour la reconnaissance de la parole, on transforme une onde parlée en une séquence de symboles (phonèmes), le synthétiseur de parole réalise la transformation inverse : à partir d'une séquence de symboles (phonèmes ou lettres), il fabrique une onde parlée semblable à celle qu'émet un locuteur humain prononçant la même séquence de sons [1].

Les domaines d'utilisation d'un synthétiseur sont multiples. Nous en donnerons les trois types principaux.

Un synthétiseur peut servir de modèle articulatoire ; en d'autres termes, il permet d'étudier la relation entre l'articulation et la production du son.

Un synthétiseur peut aussi permettre d'étudier quels sont les paramètres phonétiques qui sont importants pour la perception. Une des grandes difficultés, dans le domaine de l'analyse de la parole, est la masse de données fournies par l'onde acoustique.

Quels sont donc les paramètres de cette onde qui sont pertinents, et quels sont ceux qui peuvent être ignorés, en première approximation tout au moins ?

Un synthétiseur peut enfin être utile pour une transmission de la parole plus économique mais toujours efficace. Par la théorie de l'information, nous savons que si un canal a une largeur de bande ΔF (en Hz) et que le signal à transmettre et le bruit ont les puissances respectives S et B, il a une capacité $C = \Delta F \log_2 \left(1 + \frac{S}{B}\right)$ bits/s pour une retransmission à taux d'erreurs arbitrairement faible. Une ligne de téléphone a une largeur de bande de 3000 Hz ou plus et un rapport signal-bruit d'environ 30 dB. Le canal en question a donc une capacité minimum de 30 000 bits/s. Dans un système de modulation à impulsions codées (PCM), le signal est échantillonné au taux de Nyquist ($2\ \Delta F$) et pour maintenir la distorsion à un niveau tolérable, l'amplitude est quantifiée à 1 ou 2 % près. Donc pour une quantification à 64 niveaux (6 bits), la capacité requise est de 2.3000. $\log_2 64 = 36\,000$ bits/s. Si chaque phonème français x_i (il y a 36 phonèmes) avait la même probabilité $p(x_i)$, l'information moyenne donnée par un phonème serait :

$$H(X) = -\,p(x_i) \sum_i \log\ p(x_i) = -\,36 \cdot \frac{1}{36} \cdot$$

$$\cdot \log_2 \frac{1}{36} = \log_2\ 36 = 5{,}17 \text{ bits.}$$

[1] Si l'on imagine une machine capable de traduire la forme parlée d'une langue A en la forme parlée d'une langue B, l'onde parlée A passerait d'abord par une machine à reconnaître la parole (de la langue A), pour être ensuite traduite (en langue B) et finalement synthétisée par un synthétiseur (de langue B).

En fait, certains phonèmes sont plus fréquents que d'autres et l'information moyenne d'un phonème est légèrement inférieure à 5 bits. A un taux moyen de dix phonèmes par seconde, l'information est donc transmise à la cadence de 50 bits/s. Cela ne veut pas dire cependant que l'onde acoustique contient 600 fois plus d'information que la séquence de phonèmes, ni que l'onde acoustique est un code inefficace, ni que le cerveau humain peut traiter 30 000 bits d'information à la seconde. Intuitivement, il est évident que le signal acoustique contient plus d'information que la séquence de phonèmes (des renseignements sur le locuteur par exemple). Cependant une grande partie de cette information est redondante. Quelle proportion ? c'est ce qui est difficile à dire. Le taux d'information d'une source continue dépend du critère de fidélité adopté. Il est clair que si seule la reconnaissance de la parole est requise, la quantité d'information nécessaire sera plus faible que si l'on veut reconnaître la parole et le locuteur.

Bien qu'il ne soit vraisemblablement pas possible de dire exactement combien l'onde parlée contient d'information, des expériences avec de la parole synthétique ont montré que l'on peut retransmettre de la parole de très bonne qualité au moyen de canaux ayant une capacité bien inférieure à 30 000 bits/s. En fait, on peut abaisser cette capacité aux alentours de 1000 à 2000 bits/s.

Les synthétiseurs peuvent se diviser en trois catégories principales, selon qu'ils essaient d'imiter la production de la parole à partir d'un modèle mécanique, électrique, ou mathématique.

Les modèles mécaniques

Si l'on omet les cas des statues parlantes et des oracles « miraculeux » de l'Antiquité (où en général la voix parvient à la « bouche » par l'intermédiaire de tubes bien dissimulés), les premiers essais de fabriquer une machine qui parle remontent à la fin du XVIII^{e} siècle. Wheatstone (plus connu pour son pont de mesure) perfectionna une machine inventée par Kratzenstein et améliorée par von Kempelen. L'appareil comportait un soufflet imitant l'action des poumons, un sifflet produisant des sons assez semblables aux fricatives, une anche (semblable à celle d'un saxophone) imitant l'action des cordes vocales, et un tube résonnant de cuir grâce auquel diverses voyelles pouvaient être obtenues, selon la déformation donnée au tube. Le tout pouvait être contrôlé par une seule personne, et von Kempelen pouvait produire dix-neuf sons de façon « satisfaisante ».

Alexander Graham Bell, inventeur du téléphone et fils d'Alexander Melville Bell, phonéticien avant l'heure, construisit un modèle « physiologique », dans lequel les articulateurs (lèvres, langue, voile du palais) étaient actionnés par des fils reliés à un clavier. Bell réussit à produire des voyelles et des consonnes nasales.

D'autres tentatives ultérieures, par Riesz notamment, ont donné des résultats légèrement meilleurs. La production de sons isolés peut être remarquablement bonne, mais l'obstacle majeur est le contrôle dynamique du synthétiseur, dans la production de séquences de sons. Il n'existe pas encore de matériaux synthétiques ou naturels qui aient des propriétés mécaniques comparables à celles des muscles et des tissus vivants de l'appareil vocal humain. Cependant la construction d'un synthétiseur statique peut tout de même être d'un certain intérêt pour l'étude de problèmes particuliers, tels que l'écoulement de l'air à la glotte ou la production des fricatives par exemple.

Les modèles électriques

Ceux-ci se subdivisent en deux types : les modèles physiologico-électriques (appelés « Vocal Tract Analogs » en anglais) et les modèles acoustico-électriques (appelés « Terminal Analogs » ou encore synthétiseurs de formants). Dans un modèle du premier type, l'appareil vocal est considéré comme une succession de tubes homogènes, chacun ayant une section différente. Si les dimensions physiques d'un tel circuit acoustique sont petites par rapport à la longueur d'onde, l'analogue électrique de ce circuit est une ligne de transmission asymétrique constituée d'autant de sections de filtres (en T, en Π, ou en Γ) qu'il y a de segments de tubes (fig. 12). Si l'on connaît la section A(x) en fonction de la distance x à la glotte, les éléments L et C peuvent être calculés pour chaque segment.

Le circuit est complètement déterminé si l'on connaît l'angle de perte de chaque élément. En notation complexe, la fonction de transfert de cette ligne de transmission est de la forme

$$\frac{U_2(s)}{U_1(s)} = H(s) = \prod_k \frac{s_k \cdot \bar{s}_k}{(s - s_k)(s - \bar{s}_k)}$$

où $s = \sigma + j\omega$ est la fréquence complexe ; $U_1(s)$ est la transformée de Laplace du débit d'air, et s_k et $\bar{s}_k$ sont les pôles complexes conjugués du k-ième segment. Pour transformer cette ligne de transmission en un synthétiseur, il faut encore y ajouter trois générateurs de signaux et finalement un transducteur pour retransformer l'onde électrique en une onde acoustique. Pour la production des voyelles et des consonnes voisées, un générateur fournit des impulsion triangulaires (dont la fréquence correspond à la fréquence de vibration des cordes vocales) à l'entrée (côté glotte) de la ligne de transmission. Pour la production des fricatives, un générateur de bruit est introduit dans le segment de la ligne qui correspond au point d'articulation de la fricative en question.

De façon semblable, pour les occlusives, une impulsion isolée est produite dans le segment correspondant au point d'occlusion. Il est évident que le contrôle dynamique et la synchronisation de tous ces éléments ne peuvent se faire qu'avec des composantes R, L et C variables électroniquement, et de préférence à l'aide d'un calculateur digital. Le premier synthétiseur de ce type, construit par Dunn en 1950, était statique, ne synthétisait que des voyelles,

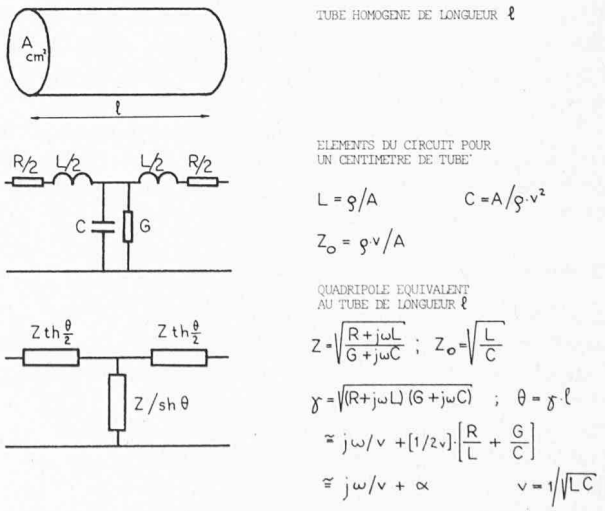

Fig. 12. — Tube de longueur arbitraire et l'un de ses circuits électriques équivalents.

et tous les contrôles devaient se faire manuellement. En 1960 Rosen, à M.I.T., a construit un synthétiseur dynamique, DAVO. Il utilisait une mémoire électronique pour le stockage des données dynamiques, ce qui limitait la longueur des segments de parole à deux ou trois sons. Un synthétiseur du même type, plus complet, est en construction à l'Université de Colombie britannique. Grâce à l'emploi de circuits intégrés, tous les circuits peuvent être logés dans une armoire métallique (environ 1 m^3), contrôlés par un ordinateur PDP-9, et programmés à partir d'un téléscripteur.

Au lieu de considérer la synthèse de la parole comme une modification continuelle d'une onde acoustique pendant sa propagation, on peut examiner le résultat global. C'est ce que font les synthétiseurs « terminaux ».

Le système le plus simple (appelé « Pattern Playback » et construit par les laboratoires Haskins à New York, vers 1950), est une sorte de spectrographe inversé. On dessine un spectrogramme plus ou moins schématique sur une bande transparente sans fin. Par balayage optique, on reconstruit le signal à partir de son spectre. La vitesse de balayage étant constante, la parole produite est monotone, ce qui est un inconvénient non négligeable. De plus l'intensité de chaque harmonique ne peut pas être reproduite très fidèlement. Un avantage de ce système, très utilisé jusqu'à très récemment, est sa simplicité d'emploi. De plus, contrairement aux autres synthétiseurs actuels, une phrase de quelques secondes peut être synthétisée en une minute ou moins par un expérimentateur exercé.

Le synthétiseur « terminal » le plus courant est le synthétiseur de formants. On en trouve deux versions principales : la version « parallèle » et la version « série ». Dans la première, chaque formant est produit séparément et contrôlé en fréquence et en intensité avant d'être combiné aux autres composantes. Dans la version « série », la sortie d'un circuit tient lieu d'entrée pour le suivant. L'intensité des formants ne peut pas être ajustée séparément, mais en fait, c'est ce qui se passe dans l'appareil vocal pour la production des voyelles orales.

Pour les voyelles nasales, pour les fricatives et pour les nasales par contre, la configuration « parallèle » semble être une meilleure approximation de la réalité. C'est ce qui a poussé Fant et Mártony, à l'Institut royal de technologie de Stockholm, à combiner les configurations parallèle et série dans leur synthétiseur OVE II (fig. 13). La qualité de la parole produite par OVE II est très bonne (au point que l'auditeur a parfois de la peine à différencier la parole synthétique de la parole naturelle) et il est possible aujourd'hui d'acheter un synthétiseur dérivé d'OVE II, EVA-MK III.

C'est même le seul synthétiseur complet que l'on trouve sur le marché. Les onze paramètres qui peuvent être ajustés pour la synthèse sont : la fréquence F_0 et l'amplitude A_0 de la fondamentale ; les fréquences F_1, F_2 et F_3 des trois premiers formants ; les amplitudes A_F, A_H et A_N des fricatives, des consonnes aspirées et des nasales respectivement ; et les fréquences K_0, K_1 et K_2 d'une antirésonance et des deux premières résonances des fricatives [1].

Chacun des onze paramètres variables est reporté en ordonnée sur une feuille de mylar, au moyen d'une encre conductrice, l'abscisse représentant le temps. Lorsque les onze têtes de lecture se déplacent (parallèlement à l'axe des abscisses), chaque ligne tracée engendre une tension électrique qui contrôle un paramètre particulier dans le synthétiseur.

Les « Vocoders »

Un « Vocoder » (Voice Coder) est un tandem *analyseur-synthétiseur* ayant un but très particulier : permettre la retransmission de la parole de façon intelligible avec un canal d'une capacité bien inférieure aux 30 000 bits/s dont nous avons parlé plus haut. Bien que des chercheurs y travaillent depuis trente ans, nous n'en avons pas parlé plus tôt, parce que le « Vocoder » évite (vu son but particulier) le problème crucial : la reconnaissance des phonèmes (ou au moins des sons) et la synthèse à partir de ces mêmes phonèmes.

[1] N_0, N_1, N_2, N_3 et N_4 représentent respectivement, une antirésonance et les quatre premières résonances nasales (toutes fixes). K_H représente une résonance fixe des consonnes aspirées.

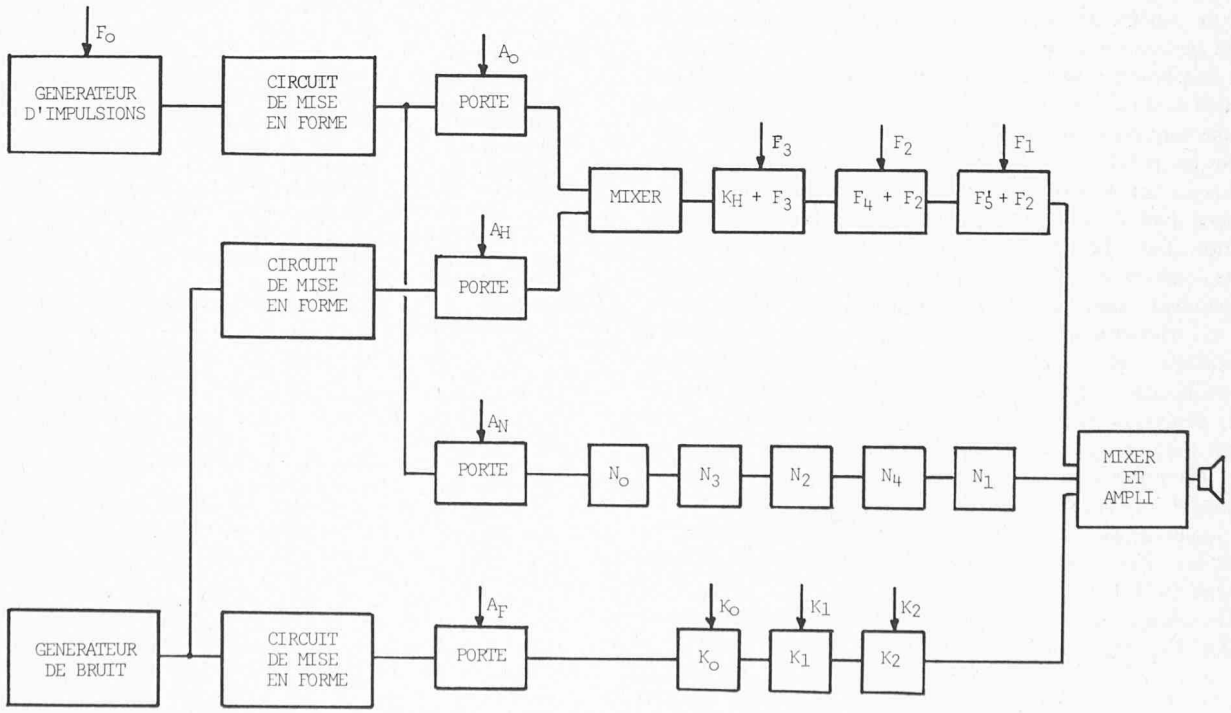

Fig. 13. — Schéma de principe de OVE II.

L'analyseur retransmet simplement, au moyen d'un système multiplexe (en fréquence ou en temps), la sortie d'une série de filtres de bande contigus étagés ; un canal supplémentaire détermine si le son est voisé ou non, et si c'est le cas, il en donne la fréquence fondamentale. Le synthétiseur reconstruit une approximation de la parole originale à partir de ces différents signaux. Les « Vocoders » ont donné des résultats relativement bons, étant donné que pour préserver une intelligibilité satisfaisante et un son naturel, il n'est pas nécessaire de reproduire en détail l'onde originale mais simplement une réplique grossière de son spectre. Il ne faut toutefois pas perdre de vue que s'ils analysent la parole, ils n'en font pas la reconnaissance. Leur qualité moyenne provient en grande partie de la difficulté de construire un bon détecteur de fondamentale.

Les modèles mathématiques

La capacité et la rapidité des ordinateurs actuels permet de reconnaître et de synthétiser la parole presque entièrement par software. Un « Vocoder » d'un type nouveau a été réalisé par Oppenheim. Il utilise avec succès le concept de *cepstrum*. Le cepstrum est par définition le spectre d'intensité du logarithme du spectre d'intensité [1].

L'analyse consiste en une mesure du cepstrum qui permet de déterminer si le son est voisé et, si oui, quelle est sa fondamentale. Lorsqu'il y a voix, il y a des maxima très marqués dans le cepstrum et ces maxima se trouvent à des *quéfrences* [1] multiples de la période de la fondamentale. L'enveloppe spectrale est caractérisée par échantillonnage du cepstrum. La synthèse s'obtient par la reconversion des résultats de l'analyse : la parole synthétique est créée par convolution de la fonction d'excitation (fondamentale) et de la fonction de transfert (de l'appareil vocal moins la glotte). Ce « Vocoder », avec un canal d'une capacité de 7800 bits/s, ne crée pas de distorsion perceptible. On peut donc supposer qu'il est possible d'abaisser encore considérablement cette capacité.

Synthèse par règles

Les synthétiseurs que nous avons examinés remplissent tous l'un des trois buts mentionnés plus haut : ils servent soit de modèle articulatoire, soit d'auxiliaire dans l'étude de la perception (pour des tests d'audition par exemple), soit de « compresseur » de largeur de bande. Aucun de ces synthétiseurs ne peut encore être programmé à partir d'une simple séquence de phonèmes. Le but de la synthèse par règles est précisément d'arriver à établir les règles phonologiques, ou plus généralement linguistiques, qui déterminent quel allophone particulier sera utilisé pour la réalisation d'un phonème dans un contexte donné. Les deux mots « déci » et « déçu », par exemple, transcrits phonémiquement /desi/ et /desy/, ont tous deux le phonème /s/ en troisième place. Cependant les deux réalisations phonétiques seront différentes : pour « déci » ce sera un [s] prononcé avec les lèvres écartées à cause du [i] qui le suit, alors que pour « déçu » le [s] sera prononcé avec les lèvres arrondies, à cause de la voyelle arrondie [y] qui le suit. Ce phénomène de coarticulation est très important en phonétique, et tout synthétiseur devra finalement inclure ces règles de coarticulation, de même que d'autres règles dont les effets portent plus loin que simplement sur les voisins immédiats.

Les ordinateurs actuels permettent au chercheur de travailler en temps réel, c'est-à-dire qu'il peut obtenir le résultat d'une modification de sa stratégie presque instantanément, alors qu'il y a quelques années seulement, des limitations de mémoire, et parfois de vitesse, rendaient pratiquement impossible une telle façon de procéder.

[1] Le domaine de résolution du cepstrum n'est pas la fréquence, mais quelque chose qui a la dimension du temps ; pour le distinguer du temps et de la fréquence, on l'a baptisé *quéfrence*.

Conclusion

Alors que la synthèse de la parole est à un stade de développement déjà avancé, la reconnaissance automatique de la parole se heurte encore à des problèmes presque insurmontables. L'une des raisons de cet état de choses est que si la transformation *configuration de l'appareil vocal – onde acoustique* est univoque (cas de la synthèse), la transformation inverse ne l'est pas. Pour déterminer cette dernière, des données considérablement plus nombreuses sont nécessaires. En fait, ce n'est que par une étude interdisciplinaire de base entreprise par des linguistes, des acousticiens, des neurophysiologistes et des ingénieurs que ce problème aura des chances d'être résolu avec succès.

Adresse de l'auteur :
André-Pierre Benguerel, Faculty of Medicine,
Division of Audiology and Speech Sciences
University of British Columbia, Vancouver 8 B.C., Canada.

BIBLIOGRAPHIE

CHOMSKY, N. A. and HALLE, M. : *Sound Patterns of English*, New York, Harper and Row, 1968.
DAVIS, K. H., BIDDULPH, R. and BALASHEK, S. : Automatic Recognition of Spoken Digits. *J. Acoust. Soc. Am.*, 24, 1962, 637-642.
DUDLEY, H. and BALASHEK, S. : Automatic Recognition of Phonetic Patterns in Speech. *J. Acoust Soc. Am.*, 30, 1958, 721-732.
DUNN, H. K. : The Calculation of Vowel Resonances and an Electrical Vocal Tract. *J. Acoust. Soc. Am.*, 22, 1950, 151-166.
FANT, G. and MÁRTONY, J. : Speech Synthesis. *KTH Stockholm, STL-QPSR* 2/1962, 18.
FLANAGAN, J. L. : *Speech Analysis, Synthesis and Perception*, Berlin, Springer-Verlag, 1965.
FRY, D. B., and DENES, P. : The Solution of Some Fundamental Problems in Mechanical Speech Recognition. *Language and Speech*, 1, 1958, 35-58.
FUJIMURA, O. : The Nagoya Group of Research on Speech Communication. *Phonetica*, 7, 1961, 160-162.
GRÜTZMACHER, M. und LOTTERMOSER, W. : Über ein Verfahren zur tragheitsfreien Aufzeichnung von Melodienkurven. *Akust. Z.*, 2, 1937, 242-248.
KEMPELEN, W. VON : *Le mécanisme de la parole*, suivi de *La Description d'une machine parlante*, Vienne, J. V. Degen, 1791.
KOENIG, W., DUNN, K. H. and LACEY, L. Y. : The Sound Spectograph. *J. Acoust. Soc. Am.*, 18, 1946, 19-49.
KRATZENSTEIN, C. G. : Sur la naissance de la formation des voyelles. *J. Phys.*, 21, 1782, 358-380.
MALMBERG, B. : *La Phonétique*, Paris, « Que sais-je », 1954.
MEYER-EPPLER, W. : *Grundlagen und Anwendungen der Informationstheorie*. Berlin, Springer-Verlag, 1959.
NOLL, A. M. : Cepstrum Pitch Determination. *J. Acoust. Soc. Am.*, 41, 1967, 293-309.
OPPENHEIM, A. V. : Speech Analysis-Synthesis System Based on Homomorphic Filtering. *J. Acoust. Soc. Am.*, 45, 1969, 458-465.
PETERSON, G. E., and SHOUP, J. E. : The Elements of an Acoustic Phonetic Theory. *J. Speech Hear. Res.*, 9, 1966, 68-99.
POSTAL, P., *Aspects of Phonological Theory*, New York, Harper and Row, 1968.
ROSEN, G. : Dynamic Analog Speech Synthesizer. *J. Acoust. Soc. Am.*, 30, 1958, 201-209.
SCHROEDER, M. R. : Recent Progress in Speech Coding at Bell Telephone Laboratories. *Proc. 3rd Congr. Int. Acoust.*, Stuttgart, 1959.
SHANNON, C. E. : Prediction and Entropy of Printed English. *Bell Syst. Tech. J.*, 30, 1959, 353-354.
SLAYMAKER, F. H., and HOUDE, R. A. : Speech Compression by Analysis-Synthesis. *J. Audio. Eng. Soc.*, 10, 1962, 144-148.
STEVENS, K. N. : Toward a Model for Speech Recognition. *J. Acoust. Soc. Am.*, 32, 1960, 47-55.
WHEATSTONE, Sir CHARLES : *The Scientific Papers of Sir Charles Wheatstone*, London, Taylor and Francis, 1879.