T. TARNOCZY

N°43. SEPTEMBRE 1969

Réflexions sur le problème de la

RECONNAISSANCE AUTOMATIQUE de la parole

suivies d'un entretien avec T. TARNOCZY

sur :

- Les vases acoustiques
- Le problème de la parole de KEMPELEN à nos jours

par E. LEIPP

G A M

BULLETIN DU GROUPE D'ACOUSTIQUE MUSICALE
FACULTÉ DES SCIENCES _ 8 RUE CUVIER. PARIS 5e

GROUPE D'ACOUSTIQUE MUSICALE
LABORATOIRE D'ACOUSTIQUE
Faculté des Sciences
8 Rue Cuvier - PARIS 5°

Adresse Postale
9, Quai Saint Bernard 5°

PARIS, le 7 Octobre 1969

BULLETIN N° 43

REUNION DU 22 SEPTEMBRE 1969

M. le Vice Doyen GAUTHIER, en voyage à l'étranger n'a pu assister à notre réunion à son grand regret, ainsi que M. le Professeur SIESTRUNCK, retenu par une réunion importante relative à nos recherches sur la synthèse de la parole.

Etaient présents :

M. LEIPP, Secrétaire général
Melle CASTELLENGO, Secrétaire

Puis, par ordre d'arrivée :

J.S. LIENARD (Ingéieur A et M) et Madame LIENARD, Professeur de musique. M. JAMET (IDN) et Mme JAMET, professeur de français; Melle HIRSCHLER (Psycho-sociologue); M. BORIS (architecte, RAUC); M. DUPREY (RAUC); M. FRANCOIS (Laboratoire d'acoustique de l'EDF); M. CARRE (R) (Ecole Nationale Supérieure d'électronique de Grenoble); M. BRAURE (CNRS); M. TEIL (CNRS, Centre de Calcul analogique); M. QUINIO (CNRS, CCA); M. RENARD (CNRS,CCA); M. NICLOUX; M. GUIBERT; M. A. TAMBA (Compositeur CNRS); M. CARCHEREUX (maître luthier); Mme de CHABURE (Conservatrice du Musée Instrumental du Conservatoire de Musique); M. DEMARAIS (Office National de Recherches Aéronautiques; M. MOUTET (Office National de recherches aéronautiques); M. MECZ (Professeur); M. FONAGY (professeur Institut de Phonétique de Paris); M. JOUHANNEAU (CNRS, Collège de France); Melle M. GUERSTEIN, Informaticienne, SEMA); M. DUBUC, Ingénieur Centre National des Arts et Métiers); Dr KADRI (orthophoniste); M. CONDAMINES (Laboratoire d'Acoustique de l'ORTF); Mme BOREL MAISONNY (Rééducation de la parole); M. CHASTE (Centre de recherche Compagnie Générale d'Electricité (CGE) et ses deux collaborateurs.

Excusés : M. J. CHAILLEY, Directeur de l'Institut de Musicologie, en voyage en URSS; M. GALLOIS-MONTBRUN, Directeur du Conservatoire National de Musique de Paris; M. J. PERROT, Directeur de l'Institut de phonétique de Paris; M. SELMER (Instruments de musique); Melle COURTIN (Inspectrice générale de Musique); M. TRAN VAN KHE (CNRS); M. MOULIN (Centre de Prospective et d'évaluation); M. BATISSIER (SUPELEC); M. [illegible] (Directeur Centre de recherches physiques de Marseille); M. BUSNEL (Directeur Laboratoire d'acoustique animale INRA); M. GARDERET (Musicologue); M. J. CAMION (architecte).

PERIODIQUE : 6 numéros annuels
Prix de vente 1 service gratuit
Imprimeur : Labo. d Mécanique Physique de la Fac. des Sciences de Paris
Nom du Directeur : M. le Professeur SIESTRUNCK

N° d'inscription à la commission paritaire : 46 283

N°43

REFLEXIONS SUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

T. TARNOCZY

La parole est une suite de signes, composée d'éléments acoustiques originaux qu'on ne retrouve dans aucun autre système de signes acoustiques. Les règles d'association de ces éléments et les "superformes" combinatoires réalisées varient selon les communautés linguistiques, de façon plus ou moins arbitraire, mais permettent à chacune d'exprimer ainsi n'importe quelle pensée.

Mais les paroles s'envolent sans laisser de trace ! C'est pourquoi l'un des plus anciens rêves de l'humanité est l'enregistrement, la reproduction et l'imitation artificielle de la parole. Entre autres problèmes techniques importants, aujourd'hui, il faut citer celui de la compréhension automatique de la parole en vue de la réalisation de machines commandées par la voix humaine.

Supposons que la parole existe depuis l'apparition de l'homme en tant qu'être intellectuel et social, donc à peu près depuis 600 000 ans. De leur côté, les hiéroglyphes permettant d'exprimer certains événements et pensées - bien sûr dans une proportion limitée - n'ont guère plus de 6000 ans. Evidemment cette invention a marqué un grand pas dans le domaine de la fixation des événements et des pensées, mais pour noter le contenu total de la parole, elle n'était pas suffisante. En effet on ne peut pas considérer les hiéroglyphes comme transcription directe de la parole, mais plutôt comme une certaine façon de communiquer des pensées. Cependant, cette forme optique, pour primitive qu'elle était, avait quand même de nombreux avantages. Elle est "matérielle" par opposition à la parole, et universellement, donc internationalement intelligible par opposition au langage.

L'imperfection des méthodes idéographiques s'explique : les signes élémentaires d'idéographie primitive ne correspondent pas aux signes élémentaires de la parole (des voix), et représentent des "unités" de langage (des concepts) beaucoup plus grandes que les lettres ou les syllabes. C'est théoriquement une économie, mais au lieu de 20 à 40 signes (lettres), il faut, dans le cas idéal 40 000 à 60 000 signes idéographiques ou signes de mots. De nos jours la branche idéographique est représentée par l'écriture chinoise.

L'écriture occidentale a suivi une autre direction. Nous n'avons pas toutes les données exactes des progrès réalisés, mais le Musée Archéologique de Beyrouth et le Musée Britannique de Londres apportent de nombreux éléments objectifs de certaines phases du progrès.

Selon notre avis, justifié par les documents des collections mentionnées, le développement de l'écriture alphabétique n'est pas la conséquence des règles intérieures d'idéographie, mais résulte d'un processus destiné à l'origine à transcrire la parole. Le problème était alors non de représenter optiquement les concepts liés aux mots, mais de procéder à un codage, correspondant au découpage acoustique de la parole matérialisée sous forme optique.

Dans l'histoire de l'humanité cette idée était extrêmement nouvelle et révolutionnaire. Pour reconnaitre les signes optiques, il faut de grandes capacités d'abstraction; la parole se compose alors de quelques types restreints des signes acoustiques, les phonèmes, qu'il faut

...../

réussir, en temps réel, à mettre en corrélation avec l'assemblage des signes optiques, des lettres. Les difficultés de ce problème se manifestent souvent : certains chercheurs linguistiques sont en train de discuter, si les affriquées sont des signes acoustiques autonomes ou composées par deux autres.

Quoi qu'il en soit, c'est sans doute depuis les phéniciens qu'existe un système composé de 22 (aujourd'hui 26) signes, avec lequel les phonèmes de la quasi totalité des langues se transcrivent plus ou moins bien. Le défaut de cette méthode est que seul l'homme ayant appris à lire peut passer indifféremment des signes écrits aux signaux acoustiques et inversement. De plus, l'écriture ne contient que peu d'éléments en ce qui concerne certaines intentions du locuteur qui veut communiquer de l'information par la parole; par exemple le timbre personnel de la voix, l'accent particulier.

Beaucoup d'entre nous pensaient que le problème posé était résolu dès l'apparition du phonographe d'Edison, puis celle des magnétophones modernes qui enregistrent et reproduisent fidèlement la qualité de la parole. La transformation puis la retransformation inverse de la parole sont bien assurées par ces machines ! Mais les choses ne sont pas si simples et le problème reste posé aux chercheurs malgré les progrès actuels de la science et de la technique.

Voyons l'ensemble de la chaine de communication humaine. Un ordre partant du cerveau de l'émetteur met en mouvement les organes vocaux pour communiquer une pensée. Celà est possible grâce à une série de signes acoustiques. L'organe de l'ouie du récepteur déforme plus ou moins les signes et les transmet au cerveau qui décode le message de l'émetteur. Le décodage se fait à divers niveaux. Le premier niveau,, le plus bas, est la compréhension phonétique; le second est la compréhension linguistique; le plus élevé est la compréhension du sens.

La liaison entre les pensées du récepteur et de l'émetteur est donc assurée par l'intermédiaire d'une série de signes acoustiques. Au point de vue de la pensée, les signes acoustiques sont codés ("boded" en anglais). C'est grâce aux particularités du cerveau humain que l'homme sait faire le codage et le décodage ("coding-decoding"). Les appareillages télétechniques, aussi bien que les appareillages d'enregistrement et de reproduction ne sont capables que de stocker ou de convoyer les signes acoustiques, de les répartir dans l'espace ou dans le temps. Mais ils ne savent pas les décoder.

Qu'en est-il de la communication humaine avec les signes optiques ? Notre cerveau sait commander des mouvements de l'appareil phonatoire pour fabriquer des signes acoustiques codés, mais il sait aussi piloter les mouvements pour former les signes optiques de l'écriture. Le code conventionnel entre les phonèmes et les caractères écrits permet la fixation écrite de la parole. Le traitement de l'information contenue dans cette série de signes optiques a lieu dès le passage dans les yeux, puis dans le cerveau. Bref, le cerveau humain peut choisir deux systèmes de signes codés; l'information est la même dans les deux cas, mais les signes sont de forme différente.

Pour la réalisation des signaux nous disposons d'organes adéquats et pour leur détection nous avons les organes sensoriels. Seul le cerveau est capable de passer du système de signes acoustiques à celui de signes optiques ou inversement. Remplacer le cerveau par des moyens technologiques pour décoder les messages parlées est l'une des tâches les plus difficiles et les plus importantes de la recherche dans le domaine de la parole. Pour la réalisation de certaines tâches relevant de la commande automatique, il faudrait éliminer de la chaine

..../

Fig. 1

TRANSCODAGE par L'HOMME

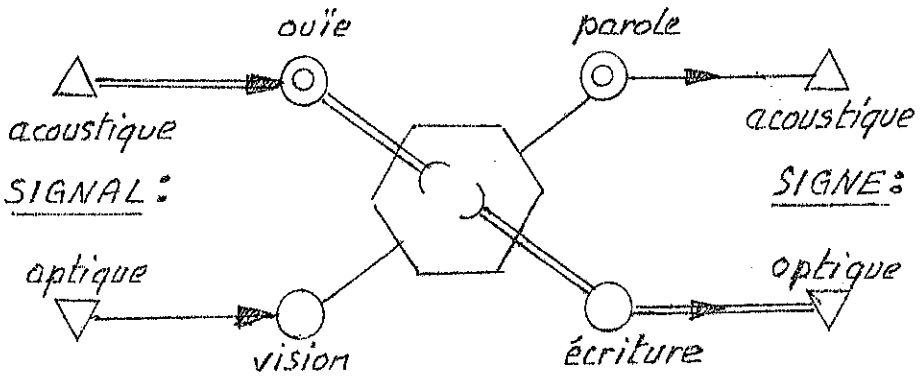


Fig. 2

TÂCHE de la MACHINE

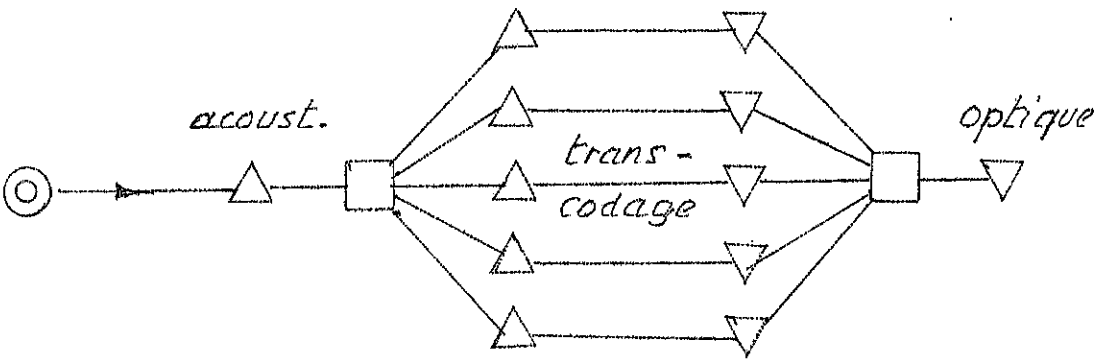


Fig. 3

le VOCODER

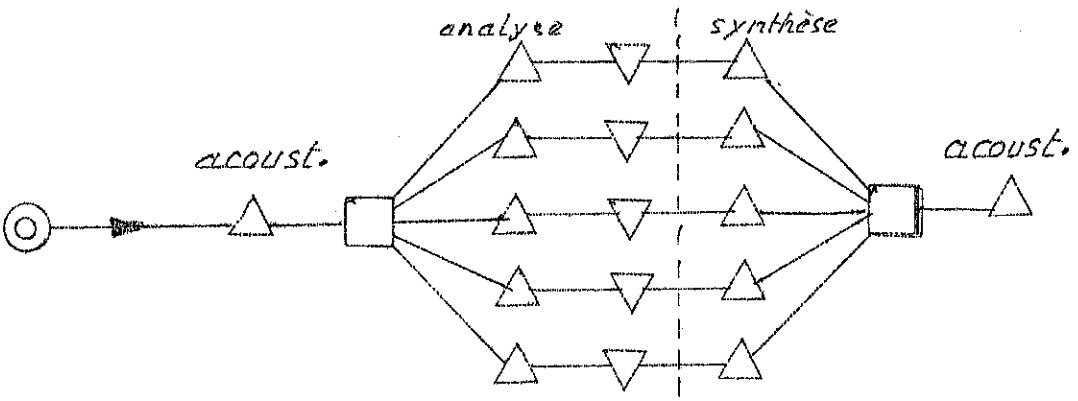
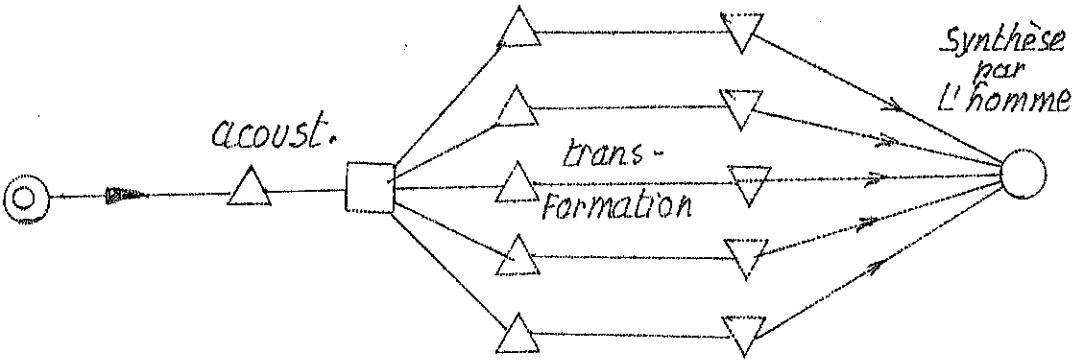


Fig. 4

VISIBLE SPEECH (Sonagraphe)



de communication l'homme considéré, d'abord comme récepteur. Quant à conférer aux machines le rôle proprement créateur du cerveau émetteur, on n'en voit pas actuellement la possibilité. Mais l'automatisation du récepteur offre déjà beaucoup de possibilités : des machines commandées par la voix révolutionneraient certainement beaucoup de problèmes techniques ! La tâche est donc de construire des machines qui comprennent le contenu des signes acoustiques.

La représentation schématisée du processus de communication humaine est donné en fig.1. Le décodage acoustique-optique c'est-à-dire la transformation de la parole entendue en signes écrits n'a pas lieu nécessairement au niveau cérébral le plus élevé : pour noter un discours on peut en rester au niveau phonétique. Le problème est de réussir à noter avec des " lettres " n'importe quelle combinaison de sons articulés; c'est le but que doivent atteindre les machines de transformation qui doivent noter exactement le texte entendu; mais il est évident que la machine n'a compris que la matière acoustique. Les caractéristiques d'une telle machine de transformation sont indiquées par la fig.2.

Les appareillages de transformation, connus jusqu'à présent, se divisent en trois grands groupes : machines parlantes, transformateurs de parole et analyseurs de parole. Les machines parlantes produisent les sons articulés artificiellement. La première machine parlante mécanique était celle de Farkas Kempelen, (1791) qui est le premier générateur manuel de parole conçu scientifiquement et correctement exécuté du point de vue technique. La machine, comportant des touches manipulées par les deux mains - donnait des sons articulés, modulés à partir d'une anche excitée par un soufflet. La machine de Kempelen n'est efficace qu'entre des mains habiles et entrainées. La reconstitution moderne qu'en a faite J.S. Lienard a montré de façon très convaincante l'intérêt de cette machine. En 1939 H. Dudley, R.R. Riesz et S.A. Watkins avaient déjà construit un modèle électroacoustique de la machine de Kempelen : le Voder (Voice Demonstration Operator). Evidemment le générateur de voix, construit de cette façon, doit être capable de former les phonèmes qui existent dans la langue considérée. Les recherches faites avec le VODER ont montré qu'il ne fallait pas "recopier" les signaux acoustiques correspondant aux mots, mais chercher à isoler et à définir les variables qui caractérisent l'identité et la différence entre les sons articulés. C'est dans ce sens que vont les recherches faites avec divers transformateurs de parole ou les systèmes Vocoder (Voice Coder).

Dans le Vocoder, l'analyse continue de la parole en temps réel par filtres fournit des données qu'on peut utiliser comme informations initiales dans le processus de recomposition de la parole. Nous avons schématisé les propriétés communes aux systèmes " Vocoder " fig.3. Le Vocoder résulte essentiellement de la connexion d'un analyseur et d'un synthétiseur. L'analyse fournit un ensemble des signes électriques (ou, si l'on veut, optiques) qui véhiculent le contenu informatif intégral du mot prononcé. Il existe actuellement plusieurs sortes de Vocoder. Dans certains types, les informations ne sont pas nécessairement représentées par une fraction de l'énergie partielle d'un mot passant dans une certaine bande de fréquence de tel filtre. L'analyse permet aussi d'isoler et de définir divers paramètres, par exemple l'intensité sonore, la hauteur de son, l'état sonore ou sourd, les régions fréquentielles caractéristiques où se situe l'énergie des fricatives, etc... Les modifications temporelles simultanées de ces paramètres permettent de définir exactement la succession des sons articulés. Une fois les variations paramétriques connues, le deuxième élément du Vocoder, le synthétiseur, assemble ces données électriques ou optiques avec préci-

...../

sion et reconstitue la suite normale des signes acoustiques. Il existe actuellement des Vocoders élaborés à 11 ou 7 canaux qui "parlent" très bien ! Les courbes portantes les informations des variations des divers paramètres sont dessinées sur rubans transparents, lus par la machine au moyen de cellules photoélectriques ou de dispositifs de mesure électrique.

Dans certain sens, le Vocoder est donc une machine parlante; mais il part de données commandées par la parole normale : ce n'est pas vraiment de la synthèse de parole ! Ni les véritables machines parlantes, ni le Voder, ni le Vocoder ne résolvent la tâche que nous avons définie plus haut.

Ce sont les méthodes de " Visible Speech " qui approchent le mieux le but que nous avons proposé parce que ces analyseurs matérialisent visuellement la répartition de l'énergie dans les bandes de fréquences en fonction du temps. Ils utilisent donc une transformation acoustique - optique (fig.4), mais les formes optiques contiennent relativement peu d'information, et celle-ci est très floue. On ne peut plus retransformer cette image en suite de signes acoustiques, sauf si on dispose d'un lecteur optique spécial. Le décryptage des données lues peut être fait par le cerveau humain, mais si on veut le réaliser par voie mécanique, c'est encore plus difficile que celui des informations qui entrent dans le deuxième élément du Vocoder. Ici encore on ne peut se passer de l'homme !

Le problème technique de reconnaissance automatique de la parole reste donc entier. Il faut souligner trois très importantes phrases dans le travail réalisé par le cerveau. D'abord, le cerveau doit comprendre un signal optique ou acoustique; parmi les nombreuses données disponibles, il doit retrouver la seule exacte, donc choisir. Deuxièmement beaucoup de propriétés du courant de parole nous échappent encore à cause de l'inadéquation de nos appareils d'analyse. Ainsi probablement le cerveau utilise-t-il aussi les dérivées, les intégrales carrées et les fonctions d'autocorrélation, de la fonction de parole. Troisièmement les images de références stockées dans le cerveau jouent, elles, un grand rôle dans la reconnaissance cérébrale. Le cerveau peut en effet les comparer avec le signe nouvellement arrivé, et après avoir apprécié les identités et les différences, il est capable de porter un jugement. N'oublions pas qu'il a tout le temps de faire ces opérations puisque l'activité cérébrale est beaucoup plus rapide que celle des machines à calculer connues de nos jours.

Pour résoudre le problème, une méthode efficace consiste à chercher à simuler l'activité du cerveau, tout en restant bien documenté sur les possibilités des machines à calculer. Peut-être faut-il simplifier la tâche pour la rendre plus facile : nous avons déjà signalé deux fois que la réduction en nombre des signes informatifs présente de grands avantages. Dans cet ordre d'idées, il est certain que le fait de diminuer le nombre d'éléments phonétiques pourrait simplifier les problèmes de la reconnaissance mécanique de la parole. L'expérience le prouve bien : quelques signes acoustiques, bien choisis, sont facilement identifiables moyennant une analyse simple; on peut donc aisément les reconnaitre par voie mécanique. Mais si on augmente le nombre de signes, la machine doit traiter de plus en plus d'informations partielles pour pouvoir porter un jugement précis. Or, les difficultés techniques et les frais augmentent rapidement avec le nombre de signes utilisés. D'autre part, si on ne peut certainement pas construire une langue avec deux signes acoustiques, 40 signes représentent déjà une immense redondance. A titre d'exemple, la langue hongroise permet théoriquement 9464 associations consonne - voyelle - consonne bien compréhen-

sibles. Mais en fait il n'y a que 732 combinaisons qui sn_tutilisées. Evidem ent en abandonnant certains éléments phonétiques, nous ne perdrions rien des possibilités linguistiques; du même coup, la technique de reconnaissance deviendrait beaucoup plus facile. Pour fixer les idées les 732 combinaisons consonne-voyelle-consonne du hongrois seraient réalisables avec beaucoup moins que 14 voyelles et 26 consonnes; il suffirait en fait de 6 voyelles et de 11 consonnes seulement.

Connaissant l'état des recherches et les possibilités techniques, on peut prédire que dans un avenir proche on réussira à construire un système phonétique artificiel et une langue artificielle. Cette langue aurait une phonétique et une logique faciles; elle serait donc facile à apprendre par tout le monde. En outre ce serait là une langue scientifique internationale, comme jadis le latin. Pour finir, on pourrait imaginer alors des machines relativement simples susceptibles de comprendre cette langue, puis des machines sachant écrire sous dictée. Beaucoup de problèmes relatifs à l'automatisation seraient alors résolus, qui préoccupent actuellement de nombreux chercheurs du monde entier.

BIBLIOGRAPHIE

H. DUDLEY, Journ. Acoust. Soc. Am. 11, 169-177 (1939)

H. DUDLEY - R.R. RIESZ - S.A. WATKINS, Journ. Frankl. Inst. 227, 739-764 (1939).

H. DUDLEY - T. TARNOCZY, Journ. Acoust. Soc. Am. 22, 151-166 (1950)

W. DE KEMPELEN, Le mécanisme de la parole, suivi d'une description de la machine parlante. Vienne, 1791.

J.S. LIENARD, Bull. 4° Conf. d'Acoustique, Budapest, 1967, N°20A 11

R.K. POTTER - G.A. KOPP - H.C. GREEN, Visible Speech, New York, 1947

T. TARNOCZY, Rapports 5° Congr. Intern. d'Acoustique, Liège, 1965, 371-387

T. TARNOCZY. Bull. 4° Conf. d'Acoustique, Budapest, 1967, N° 20A 2.

ENTRETIEN avec M. TARNOCZY

par E. LEIPP

M. LEIPP. M. TARNOCZY vient de nous faire un exposé sur l'état actuel de la recherche acoustique en Hongrie, et plus spécialement sur le problème de la reconnaissance automatique de la parole qui l'intéresse tout particulièrement. Je dois préciser que M. TARNOCZY dirige le groupe de Recherche Acoustique à l'ACADEMIE DES SCIENCES de BUDAPEST, qui est en fait un institut de recherche; d'autre part il est professeur d'acoustique à l'Université, (unité d'enseignement uniquement).

Les sujets qui préoccupent M. TARNOCZY sont exactement les nôtres, à savoir la structure physique des signaux acoustiques et la façon dont notre cerveau traite l'information qu'ils contiennent. Il y a cependant, vous le savez, une différence notable dans nos façons d'aborder ces problèmes. Nos recherches au laboratoire sont nettement expérimentales et notre souci primordial est de comprendre ce que font les praticiens qui manipulent continuellement les sons, généralement sans souci des théories en cours. Nous pensons que, parmi eux, les artistes du son, les musiciens ont de la chose auditive une expérience, empirique certes, mais irremplaçable. C'est pourquoi nous étudions ce que font les fabricants d'instruments, les musiciens exécutants, les compositeurs, les chanteurs etc... C'est à partir de ces recherches que nous tentons de repenser les problèmes de l'acoustique, considérée comme science des sons, y compris celle de leur intégration par l'homme. Bref, notre acoustique musicale, c'est l'acoustique tout court; mais notre approche des problèmes n'est généralement pas classique, ce qui s'est avéré fructueux de nombreuses fois déjà, et justement c'est le cas pour les problèmes de la parole.

Je voudrais demander à M. TARNOCZY s'il existe en HONGRIE des personnes qui " font " de l'acoustique musicale de cette espèce.

M. TARNOCZY. Non. Nous avions chez nous autrefois un ingénieur du nom de TAMAS TARFAS, musicien, et qui se passionnait pour les questions de théorie musicale, gammes et autres; mais il est mort. Cependant nous sommes en train de former des jeunes dans cette voie.

M. LEIPP. Ils verront, comme nous combien l'acoustique musicale est un domaine passionnant mais difficile ! Nous sommes tout prêts à vous communiquer ce que nous savons et qui représente déjà une longue expérience.

Vous nous avez parlé de beaucoup de choses de grand intérêt mais j'aimerais que vous nous disiez aussi quelques mots sur un sujet que vous avez, je crois, fortement approfondi. C'est l'affaire des vases acoustiques de l'Antiquité et du Moyen-Age. Au Colloque sur le bruit dans les Habitations (Groupement des Acousticiens de Langue Française, Marseille, Septembre 1963), nous avions présenté deux communications un peu hors série. De mon côté j'avais parlé de "La musique considérée comme bruit dans les locaux à usage d'habitation" et vous aviez rédigé un travail "SUR LES VASES DE VITRUVE", qui m'avait beaucoup intéressé.

...../

Entretemps, une thèse a été faite sur ce thème au Centre de Recherches Physiques de MARSEILLE (M. FLORIOT, 1964). Je résume brièvement la question.

VITRUVE, dans son traité(Livre V, chapitre 5) parle longuement de vases d'airain que l'on disposait dans les théâtres grecs et romains, faits "en rapport avec la grandeur des théâtres", et fabriqués de façons" qu'ils rendent, quand on les frappe, l'un le son de la quarte, l'autre le son de la quinte"... "Au moyen de cette disposition, la voix,qui viendra de la scène comme d'un centre, s'étendra en rond, frappera dans les cavités des vases, et en sera rendue plus forte et plus claire...".

Bref, selon les témoignages, ces vases jouaient manifestement un rôle de "résonateur", "amplifiant" le son. Or il est évident que sans apport d'énergie, on ne peut amplifier une onde acoustique dans le sens propre du terme.. Vous aviez alors émis l'idée qu'il s'agissait d'une espèce d'effet de HAAS, c'est-à-dire d'un tout petit décalage, d'un "écho", renvoyé par le vase, qui n'est pas perçu en tant que tel, mais qui fusionne avec le son incident pour le prolonger : d'où la sensation de plus forte intensité, alors qu'il s'agit en fait d'une réverbération.

Avez-vous poursuivi vos recherches sur ce sujet ?

M. TARNOCZY. Oui. Nous avons fait des recherches in situ et des mesures en laboratoire. On peut résumer nos résultats ainsi. Tout vase est un résonateur de HELMHOLTZ qui possède une certaine "durée de réverbération". Quand la réverbération de la salle où est placé le vase est plus longue que celle du vase, celui-ci joue le rôle d'un absorbant d'énergie acoustique; mais quand on est en plein air où il n'y a pratiquement aucune réverbération (cas des théâtres antiques), la réverbération des vases est perçue; d'où leur usage à l'époque. Ces vases provoquaient avant la lettre une véritable réverbération artificielle.

M. LEIPP. Le problème ne semble pas entièrement résolu, et nous avons dans nos projets une étude sur ces questions, dès que no s disposerons de nos nouveaux locaux (Octobre 69) comportant une salle sourde.

Mais une question assez curieuse se pose. Au Moyen Age, les architectes ont très souvent incorporé dans les murs et voûtes de leurs églises, disposés de façon apparemment incompréhensible, des vases en terre dont l'orifice seul apparait. On connait en Europe près de 200 églises où de tels vases, parfois très nombreux, existent. Il s'agissait certainement d'une tradition ayant eu autrefois un sens et une efficacité. Cette tradition, perdue comme tant d'autres, puis recopiée maladroitement, perdit alors sa signification et fut abandonnée par la suite. Or si j'en crois ce qu'on raconte à ce sujet, ces vases auraient été non des résonateurs destinés à amplifier le son, mais au contraire des "pièges à son" pour étouffer les échos indésirables. Il y a là un paradoxe : d'une part les vases amplifient, d'autre part ils étouffent... Je crois que le problème des résonateurs de ce genre est très compliqué et qu'on nc possède à leur sujet que des hypothèses; d'où l'intérêt de reprendre la question. Je pense surtout que le vase-résonateur a des propriétés définissables en tant que système isolé; mais celles-ci changent lorsque le vase est placé

...../

dans tel contexte matériel donné. Il faut alors le considérer comme un paramètre dans un ensemble complexe où les variables réagissent différemment les unes sur les autres selon le cas. De toutes façons, un problème de psycho-physiologie de la perception des sons est lié à la question; on sait donc d'avance qu'elle est compliquée. Mais c'est pour celà aussi qu'elle est intéressante et nous y reviendrons sûrement un jour.

J'ai une autre question à poser à M. TARNOCZY :

Vous avez montré que le langage parlé utilise un certain nombre de sons, de "phonèmes", 5,6 ...10 voyelles; 15, ...20 consonnes. La combinatoire des sons vocaux réalisés avec ces éléments est énorme. Si on a 50 sons élémentaires, on peut en faire 50 x 50 combinaisons, soit 2500 "mots" élémentaires différents. Vous avez alors émis l'idée que dans ces conditions, l'on pourrait faire une langue artificielle tout à fait viable et intelligible avec un nombre beaucoup plus restreint de voyelles et de consonnes. Si l'on se contentait de 15 sons élémentaires, on n'aurait plus que 225 combinaisons. L'intérêt de cette langue artificielle serait de pouvoir engager plus facilement le "dialogue avec les ordinateurs", problème dont beaucoup de chercheurs se préoccupent actuellement. Comme vous avez cité KEMPELEN dans votre exposé, cela me suggère une idée curieuse. KEMPELEN a réalisé en 1791 une machine parlante qui a fait l'objet d'une réunion du GAM (La machine parlante de KEMPELEN; par J.S. LIENARD; Bulletin GAM N° 34). J.S. LIENARD a reconstruit cette machine d'après le livre de KEMPELEN lui-même, qui en donne une description très précise et détaillée. Or cette machine, justement, ne permet de produire qu'un nombre restreint de "voyelles" et de consonnes. Avec un apprentissage suffisant du "jeu" de l'appareil et aussi un "recodage" de l'audition, on arrive cependant à parler ainsi avec des signaux acoustiques assez simples et de nombre limité. Je crois que J.S. LIENARD ferait bien de nous montrer cette machine, de nous en dire un mot et de nous en faire une petite déonstration pour ceux qui ne sont pas venus à notre réunion GAM.

J.S. LIENARD. Voilà ma reconstitution conforme à la description de KEMPELEN. Je viens de voir et de photographier au DEUTSCHES MUSEUM de MUNICH, une des machines faites par KEMPELEN lui-même . Ce dernier modèle comporte quelques variantes, en particulier une "pédale" supplémentaire qui agissa t sur la rasette de l'anche, permet de moduler la hauteur. Mais cette machine n'est pas en état de fonctionnement; je vous présente donc la mienne.

KEMPELEN a tout simplement simulé de façon grossière l'appareil phonatoire humain : d'abord un soufflet représente les poumons. Une boite à vent, à la sortie des "poumons", contient l'anche en ivoire recouverte de peau et qui simule les cordes vocales. A la boite sont connectés d'abord une espèce d'entonnoir en cuir (ou caoutchouc) déformable, qui constitue un "résonateur buccal", dont on peut varier le "formant" en le recouvrant plus ou moins. Une "sortie" de la boite fournit un son "ch"; une autre un son "ss", enfin deux "trous de nez" peuvent être obturé par les doigts en cas de besoin, pour faire des plosives par exemple.

Telle qu'elle est la machine permet, pour peu qu'on soit entrainé (et KEMPELEN l'était puisqu'il a travaillé plus de 20 ans sur cette question) d'articulier des mots ou des phrases. En voici quelques échantillons (papa,maman, ah! là là ! etc...)

...../

M. LEIPP. TARNOCZY a insisté à juste titre tout à l'heure sur l'importance de la machine à parler de ce chercheur HONGROIS... Ce n'est pas tellement le "jouet" qu'il faut y voir; mais les idées directrices qui ont conduit KEMPELEN à sa réalisation et qui ont une consonance souvent très moderne. En particulier il avait très bien compris le rôle de la prévisibilité dans l'intelligibilité de la parole, idée que nous avons développée lors du GAM sur ce thème. (Bulletin n° 37).

M. J.S. LIENARD. Oui. L'ouvrage de KEMPELEN est un véritable traité de phonation, à une époque où cette science était pratiquement inexistante. Il a fait des observations sur les mécanismes de la génération de la voix et aussi sur les processus mentaux qui président à l'intération de la parole par l'homme. En particulier, il avait suggestionner son auditoire, qui, dès lors, avec quelques bribes sonores bien placées, comprenait parfaitement les mots. L'expérience est facile à faire : écoutez cette petite phrase (échantillon à la machine de KEMPELEN) Vous n'avez rien compris ? Eh bien, j'ai dit "ça va ? "Si on fabrique encore le même signal acoustique , il est devenu tout à fait intelligible, si vous avez été prévenus de ce que celà veut dire. KEMPELEN est explicite sur certains points : sa machine ne peut faire un "t" ou un "d" ? Qu'à celà ne tienne; on les remplacera par un "p"; dans le contexte celà passe très bien.

M. LEIPP. C'est bien justement dans ce sens qu'on peut penser à un langage synthétique simplifié. Mais on peut aussi se demander pourquoi les langues humaines ne se sont pas établies sur une telle base. En réalité, ce qui peut apparaître comme redondance dans celles-ci, est certainement justifié en pratique, car il n'est pas d'exemple d'une activité humaine où l'homme ait consenti à faire des efforts inutiles pendant des millénaires ! Il est certain que les tentatives de recherche d'une langue simplifiée ne sont justifiées que par la nécessité de plus en plus pressante de trouver un moyen de communiquer oralement avec des ordinateurs. Bref, le problème n'est pas tellement d'imaginer une langue synthétique, mais de trouver un moyen de la faire reconnaître par une machine. Or, c'est là le hic ! On peut considérer actuellement le problème de la synthèse de la parole comme pratiquement résolu; celui qui ne l'est pas, et de loin, est précisément celui de la reconnaissance de la parole par l'homme. Si on savait comment fonctionne notre cerveau, ce serait beaucoup plus facile; malheureusement on en est strictement réduit aux hypothèses en ce domaine qui nous concerne tous au plus haut point pourtant... Il me semble évident que le cerveau travaille par des voies économiques que nous ignorons totalement; il n'est que de considérer son volume et les performances réalisées en tant que machine de gestion et machine de traitement de l'information auditive, visuelle etc... Il est certainement illusoire de vouloir considérer le cerveau comme une machine exclusivement digitale; c'est un centre de calcul hybride, analogue-digital, qui permet de considérables économies en particulier en ce qui concerne les mémoires, où la méthode analogique est certainement la plus économique, d'autant plus qu'il suffit de stocker non pas l'intégralité des signaux informatifs, mais les invariants qui permettent de les distinguer entre eux. Le vrai problème, brûlant celui-là, est, à mon sens, celui de la simulation des fonctions du cerveau. Nous en avons parlé de nombreuses fois déjà : on n'en sortira pas autrement et il faudra nécessairement passer par une association

...../

analogue-digital et par des processus de corrélation et d'autocorrélation.

M. TARNOCZY. Je suis, comme vous convaincu que notre cerveau travaille autrement que par méthode purement digitale. En tout cas, nous avons étudié déjà des méthodes et construit des appareillages pour reconnaitre les voyelles hongroises, à partir du triangle des formants. On sait de quoi il s'agit : sur un diagramme, on porte en abscisse la hauteur du premier formant et en ordonnée la hauteur du second. Chaque voyelle recouvre alors une certaine zone dans ce diagramme. Quand il s'agit de distinguer 3 voyelles très différentes, ou, a et i par exemple, c'est bien facile avec ce diagramme. Mais les voyelles intermédiaires se chevauchent parfois et audessus de 5 voyelles la machine a du mal à reconnaitre à coup sûr les voyelles voisines....

M. LEIPP. Nous n'en sommes pas surpris et nous avons bien étudié cette question dont nous parlons dans le bulletin sur l'Intelligibilité de la parole (n°37). Il faut cependant rajouter que deux formants sont insuffisants pour décrire la forme sémantique d'un mot. D'autre part, si on change de locuteur (homme, femme, enfant etc..) il est impossible de définir un formant par sa fréquence. Le formant 3 de tel locuteur est à la même fréquence que le formant 1 de tel autre ! Nous avons fait des expériences avec l'ICOPHONE et la voix synthétique qui sont tout à fait démonstratives à cet égard. Je ne crois donc pas qu'on puisse facilement réussir à reconnaitre une voix quelconque si on pense au triangle des formants. Notre idée est que les mots sont des "formes", des "gestalts indépendantes de la fréquence; plus exactement ce sont les rapports de fréquence et les rapports de temps qui déterminent les formes véhiculant l'information sémantique de la parole. Notre programme de reconnaissance automatique de la parole est donc construit sur d'autres bases que le vôtre. Vous partez de l'idée qu'il est intéressant de simplifier le problème en supprimant un certain nombre de "phonèmes"; nous pensons qu'il faut simplifier le problème en prenant tous les phonèmes et toutes leurs combinaisons usuelles, mais en en ramenant la forme acoustique à ce qui est juste nécessaire et suffisant pour qu'elle soit reconnue sans ambiguïté (phonatomes de notre méthode de synthèse). Il est certain que nous n'en sortirons pas sans avoir résolu divers problèmes épineux, mais solubles si nous en avons les moyens techniques : anamorphoses, corrélation et autocorrélation. Finalement, pour nous, le problème de reconnaissance automatique de la parole est un cas particulier du problème très général de la reconnaissance des formes, qu'elles soient acoustiques, optiques ou autres.

M. TARNOCZY. Je crois plutôt que le problème de la forme est une partie de celui de la parole. Si vous prenez un oscillogramme d'un mot, il est facile de le lire, d'en faire une analyse de Fourier.

M. LEIPP. Je crois surtout qu'il faudrait se mettre d'accord sur la signification du mot "forme". Nous l'entendons dans le sens des gestaltistes et considérons que le monde extérieur nous est accessible sous l'aspect de formes acoustiques, optiques et autres, perçues comme un tout, ou la sensation globale ne peut être déduite des sensations fournies par les éléments isolés. Ceci pose un problème difficile de méthode de recherche où, en acoustique, on cherche dans les méthodes classiques à décomposer un phénomène complexe en ses éléments pensant mieux pouvoir "séparer les dif-

...../

ficultés". En fait on bute alors à de grosses difficultés pour interpréter les documents physiques et les transposer dans le domaine psycho-physiologique

D'autre part, je crois quand même devoir insister encore une fois sur des critiques que j'ai maintes fois formulées au sujet de l'oscillographe lorsqu'il s'agit de sons musicaux et de parole. La parole s'est élaborée au cours des millénaires en adaptation réciproque avec l'audition. L'information sémantique de la parole est concentrée au mieux autour du maximum de sensibilité de l'oreille (vers 1500 Hz), plus qu'autour du "fondamental" de la voix (autour de 100 Hz). Or l'oscillogramme qui donne l'évolution des amplitudes en fonction du temps, met surtout en valeur les phénomènes graves, de grande amplitude alors qu'elle dissimule sous la largeur du trait des phénomènes plus aigus, de très faible amplitude souvent, mais perceptivement d'importance déterminante en ce qui concerne la forme sémantique de la parole.

D'autre part, notre système auditif semble conçu pour traiter une quantité considérable d'information acoustique considérée comme variations de fréquence dans le temps; il est beaucoup plus grossier et plus "inerte" pour l'information donnée par les variations d'amplitude ou de niveau dans le temps. C'est pourquoi nous préférons de loin traiter nos problèmes avec le visible-speech (sonagraphe) plutôt qu'avec l'oscillographe le plus perfectionné et le plus précis dont nous disposions (Tektronix). Nous savons par expérience que les diagrammes fréquence-temps peuvent être décodés visuellement pour retrouver la forme des mots, les reconnaître, les relire. Or il est totalement impossible de retrouver un mot à partir d'un oscillogramme. Mme BOREL MAISONNY, qui a beaucoup travaillé avec l'oscillographe, pourrait nous le confirmer....

Mme BOREL MAISONNY. Oui. L'oscillographe est précieux pour étudier le rythme et la durée; mais en effet on ne peut "relire" un oscillogramme si on ne sait pas ce qui est dit.

M. LEIPP. J'en conclus précisément que nos sens qui fonctionnent probablement tous de la même façon quant au traitement de l'information s'intéressent surtout à l'évolution temporelle des composantes spectrales des sons. C'est pourquoi, avec de l'entrainement, on peut relire un sonagramme. Il existe d'ailleurs une preuve supplémentaire à ce que je dis. L'oscillographe tient compte de la phase, qui détermine grandement l'aspect visuel de la forme oscillographique. Or celle-ci est totalement liée à la phase. J'ai déjà cité ailleurs les expériences que j'ai faites naguère sur le violon : on joue un "sol" à vide tiré et on photographie l'oscillogramme. Puis on pousse la même note : l'oscillogramme est complètement modifié, retourné de 180°, et la forme de la courbe est fortement modifiée dans le détail. Cependant, le musicien n'y distinguera aucune différence auditive. Bien sûr que la phase existe et qu'elle intervient; mais dans un son réel, comportant 30 ou 40 composantes qui évoluent constamment et de façon plus ou moins autonome, où donc réside la "phase" du phénomène ? Si telle composante est décalée ou disparait même, celà n'altère en rien la sensation globale, tout simplement parce que la forme sémantique reste totalement reconnaissable. Si sur une photographie, mon chien a déplacé sa patte, celà ne m'empêche pas de le reconnaitre; or c'est de celà qu'il s'agit.

...../

Finalement on en revient toujours au même problème : ce qu'il faudrait, c'est savoir comment fonctionne notre système auditif, et plus particulièrement le "centre" qui traite l'information venant du "microphone-oreille". Dès qu'on le saura, qu'on en aura au moins une hypothèse de fonctionnement vraisemblable, il deviendra facile de passer à la reconnaissance automatique de la parole. C'est parce que nous sommes conscients de ce fait que nous faisons des efforts pour mettre au point le modèle de fonctionnement dont nous avons parlé déjà à plusieurs reprises. (Bulletin Intelligibilité de la parole n° 37, en particulier.)

Mme BOREL MAISONNY. Quel est l'intérêt des langages artificiels sur lesquels tout le monde travaille tant ?

M. FONAGY. Il est énorme dans la mesure où d'aucuns pensent que c'est le seul moyen pour "engager le dialogue " avec les ordinateurs.

M. LEIPP. Je crois que les personnes du Centre de Calcul Analogique du C.N.R.S. ici présents, à savoir MM. RENARD, QUINIO et TEIL qui ont mis notre méthode de synthèse de parole en machine sont bien placés pour savoir combien il serait agréable de programmer une machine par la parole plutôt que d'utiliser les moyens actuels....

De toutes façons, je remercie bien vivement M. TARNOCZY d'avoir, par son exposé, stimulé cette discussion. Les questions soulevées intéressent beaucoup de gens actuellement et je dois dire que la connaissance automatique de la parole à partir de ce que nous savons maintenant sur sa structure physique et perceptive est à notre programme. Nous espérons aboutir en deux ou trois ans à des résultats comparables à ce que nous avons obtenu en synthèse, grâce aux simplifications que nous introduisons au départ. Notre but lointain est de pouvoir reconnaitre automatiquement un discours, une phrase prononcée de façon quelconque par n'importe quel locuteur : programme ambitieux pour nous. Mais on sait bien " qu'il n'est pas nécessaire d'espérer pour entreprendre, ni de réussir pour persévérer ".....