NEUVIEME COLLOQUE SUR LE TRAITEMENT DU
SIGNAL ET SES APPLICATIONS

NICE du 16 au 20 MAI 1983

RECONNAISSANCE DE PAROLE CONTINUE MULTILOCUTEUR *

SPEAKER INDEPENDANT CONTINUOUS SPEECH RECOGNITION

P. ALINAT - E. GALLAIS

THOMSON-C.S.F.-DASM - Chemin des Travails - CAGNES SUR MER - 06801 CEDEX - FRANCE

RESUME

La reconnaissance de la parole donne lieu, en schématisant, à deux familles de systèmes : globaux et analytiques. Les premiers sont les plus répandus en raison de leur simplicité mais ne faisant que peu intervenir le fait qu'il s'agit de parole, leur utilisation est limitée par un usage contraignant. Au contraire, les systèmes totalement analytiques prennent en compte beaucoup de connaissance sur la parole. Ils sont plus compliqués et moins répandus, mais permettent un usage multilocuteur et une prononciation naturelle avec coarticulation.

L'organisation complexe d'un système totalement analytique est décrite ici : Analyseur (Cochlée artificielle) - Extraction des paramètres acoustiques - Localisation et estimation des paramètres des phonèmes - Comparaison phonèmes reconnus ↔ phonèmes vocabulaires (en utilisant les règles de coarticulation et de classification des phonèmes) - Création des hypothèses au niveau mot.

Les performances obtenues pour la reconnaissance de mots enchaînés sont données aux niveaux mots et phrases.

* Cette étude a été partiellement financée par la DRET (Contrats n° 80/048 et n° 82/047

SUMMARY

Sold speech recognition systems have a limited use because they don't take into account all special characteristics of speech : need for speaker training, restricted vocabularies, noise sensibility.

Presently studied systems are analytical, that is to say they use the known levels of speech : feature - phoneme - word. They are more complex but put less contraints on the user.

In this paper, an analytical system is briefly described, more particularly the analyser and the feature and phoneme levels. Some results are given.

RECONNAISSANCE DE PAROLE CONTINUE MULTILOCUTEUR

SPEAKER INDEPENDANT CONTINUOUS SPEECH RECOGNITION

1 - INTRODUCTION

Depuis quelques années, de nombreux systèmes de reconnaissance de parole sont en vente. La recherche dans le domaine demeure toutefois plus active que jamais parce que, mis à part quelques applications particulières, les systèmes vendus ne rendent pas les services auxquels on peut s'attendre [1] car ils comportent de nombreuses contraintes : phase d'apprentissage pour chaque nouveau locuteur, parole non naturelle (mots isolés ou enchaînés), grande sensibilité aux bruits d'ambiance. En un mot, les performances des systèmes vendus sont très pauvres par rapport à celles des auditeurs humains dans les mêmes conditions. La raison en est que ces systèmes effectuent une reconnaissance de type globale qui n'utilise qu'un module de signal à reconnaître trop général : autrement dit, pour ces sytèmes peu importe s'il s'agit de reconnaître des cris d'animaux, des chants d'oiseaux ou de la parole. Ces systèmes globaux présentent donc l'avantage de la simplicité mais ils ne constituent pas une solution satisfaisante du problème.

Pour disposer de systèmes plus performants, il faut utiliser une méthode analytique, c'est à dire prenant en compte un modèle relativement élaboré de la parole humaine. Ces méthodes conduisent à des solutions bien plus complexes que celles des systèmes globaux ce qui explique que, bien qu'un bon nombre d'équipes y travaillent depuis une quinzaine d'années, il n'y ait pas encore en vente de système analytique. Le présent exposé est relatif à la façon d'améliorer les performances des systèmes analytiques. Il porte sur les points suivants :

- Description sommaire de l'organisation couramment utilisée pour les systèmes analytiques.
- L'analyseur, souvent négligé, est une partie importante qu'il faut concevoir avec soin.
- La comparaison avec les mots vocabulaire doit être faite au niveau traits pour prendre en compte les coarticulations entre phonèmes plutôt qu'au niveau phonèmes comme cela se fait.
- Un système tenant compte de ces points a été réalisé. Ses performances sont décrites.

2 - LES SYSTEMES ANALYTIQUES

Il s'agit de systèmes utilisant plus ou moins les connaissances disponibles sur l'organisation de la parole, connaissances qui constituent la phonétique. Ces systèmes sont dits analytiques car ils tiennent compte de ce que la parole comporte plusieurs niveaux : phrase - mots - syllabes - phonèmes - traits. La compréhension de la parole par un auditeur humain utilise ces niveaux. Les objets situés à chaque niveau dépendent de la langue utilisée et ne sont pas définis avec précision : par exemple, en français, on peut considérer qu'il y a 3 ou 4 voyelles nasales.

Les systèmes analytiques étudiés ou en cours d'étude sont nombreux. On peut citer parmi les plus connus : KEAL [2], HEARSAY II et HARPY, HWIM et le système de SDC [3]. L'organisation est, en gros, la suivante :

- Un analyseur traite le signal d'entrée. Il est généralement basé sur le LPC ou un banc de filtre de VOCODEUR, c'est à dire des outils réalisés essentiellement pour réduire le débit d'information, donc dans un but de transmission.

- Une segmentation temporelle est effectuée, c'est à dire que le signal est découpé en une suite de segments dont chacun est sensé représenter un phonème ou une portion de phonème. Pour certains systèmes, cette segmentation ne fait intervenir que les maxima d'instabilité du signal, pour d'autres plus élaborées, elle fait intervenir le rythme syllabique et certains traits des phonèmes : il s'agit plutôt alors de localisation.

- Pour chaque segment, une décision est prise, c'est à dire qu'un nom est attribué : par exemple, tel segment est classé /OU/. Dans certains systèmes, plusieurs noms rangés par ordre de probabilité décroissant peuvent être attribués à un même segment. Ces noms sont obtenus soit par corrélation avec des copies (template matching), soit par des proccessus plus complexes faisant intervenir plus ou moins le niveau trait.

- La suite des segments est comparée au vocabulaire-syntaxe par des processus relativement élaborés.

Le processus n'est pas forcément montant : il peut y avoir des retours en arrière.

Les performances de ces systèmes ne sont pas encore suffisantes : apprentissage plus ou moins nécessaire pour chaque nouveau locuteur, sensibilité au bruit d'ambiance, difficulté avec de la parole continue rapide. L'impression d'ensemble offerte par ces systèmes est qu'il y a une disproportion importante entre les efforts faits aux différents niveaux : l'analyseur et l'obtention de la chaîne phonémique sont négligés par rapport aux efforts (par ailleurs nécessaires) faits pour utiliser au mieux le lexique et la syntaxe.

L'objet du présent exposé porte sur des améliorations des niveaux inférieurs : analyseur et obtention de la chaîne phonémique.

3 - L'ANALYSEUR

Il s'agit d'un dispositif qui réalise plus ou moins une analyse spectrale. Dans le cas de la transmission de Parole (Analyse - Synthèse), le choix précis des paramètres de l'analyseur n'est pas critique (bien qu'il soit tout de même sensible [4]). On peut dire que des modifications de paramètres se traduisent par plus ou moins de bruit et il est bien connu que l'auditeur humain (situé, dans ce cas, en fin de chaîne) représente un système de compréhension particulièrement robuste vis à vis du bruit dans la mesure, bien entendu, où, malgré la réduction, l'information en sortie de l'analyseur reste très redondante par rapport à l'information effectivement transmise à l'auditeur. Dans le cas de la reconnaissance, le choix a plus d'importance car les systèmes de compréhension installés derrière sont encore très primitifs comparés à un auditeur humain et de ce fait malheureusement très sensible à tout rajout de bruit. De plus, la complexité de certaines règles de classification dépend de l'analyseur et il faut donc rechercher un compromis entre la complexité de l'analyseur et celle du reste du système de reconnaissance.

Une idée simple et naturelle a donné lieu à beaucoup de travaux : il s'agit de choisir les paramètres de l'analyseur en fonction de ce qui est connu de l'oreille humaine (cochlée) [5] qui est le récepteur naturel de la parole (au cours de l'évolution, l'oreille a existé bien avant le système d'élocution humaine). A THOMSON-CSF-DASM, un tel analyseur a été réalisé et utilisé dès 1970 [6]. Il ne s'agit que d'un compromis entre la fidélité du modèle et sa complexité mais cet analyseur a tout de même permis de mettre en évidence certaines règles de classification très simples qui n'apparaissaient pas avec des moyens d'analyse classique (banc de filtre vocodeur essentiellement à l'époque) [6].

a) Pour la classification de la place d'articulation des consonnes explosives [7], les règles sont relatives à la forme spectrale du burst :

- Pour /KG/ : un pic étroit, près du F2 de la voyelle suivante (et si pas de voyelle, pas trop haut en fréquence).
- Pour /TD/ : un maximum étalé et haute fréquence ou un pic étroit au-dessus du F2 de la voyelle adjacente (haut en fréquence si pas de voyelle).
- Pour /PB/, un maximum étalé et basse fréquence ou un pic étroit au-dessous du F2 de la voyelle adjacente.

Depuis, ces règles ont été confirmées par des études plus récentes [8, 9, 10]

b) Pour la classification des voyelles et des consonnes fricatives, l'utilisation de la cochlée artificielle a permis de mettre en évidence la notion de zones formantiques [6] très utilisée dans notre système de reconnaissance.

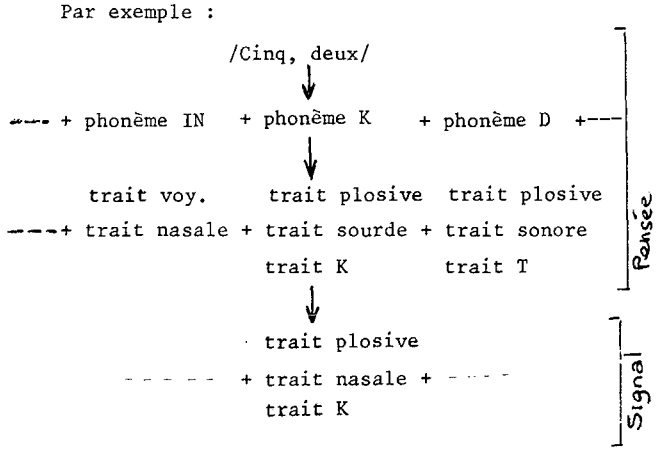
La supériorité des modèles de cochlée sur les bancs de filtres vocodeur et le LPC tient aux différences suivantes :

- L'échelle de fréquence permet une équirépartition des informations utiles (les zones formantiques en particulier, ont une longueur constante). Le LPC correspond à une échelle très différente.
- La sélectivité (fonction de la fréquence centrale) permet de faire ressortir les formants utiles à la reconnaissance. De plus, elle correspond à des constantes de temps de montée qui sont mieux adaptées à la classification des plosives.

4 - OBTENTION DE LA CHAINE PHONEMIQUE

Les systèmes analytiques classiques (voir paragraphe 2) utilisent une segmentation temporelle du signal, suivie d'une classification globale de chaque segment. Généralement, cette classification se situe à peu près au niveau phonème. A ce stade de la reconnaissance, on obtient donc une suite de phonèmes décrits par leur nom (en tolérant quelques hésitations). Par exemple, pour le mot /5,2/→ S AN N D E F
IN L B
C

C'est cette suite de noms de phonème qui va être utilisée pour la reconnaissance des mots. Cette façon de faire présente l'avantage de la simplicité. En revanche, elle utilise très mal le niveau trait. En effet, il est connu depuis longtemps, grâce notamment aux travaux des HASKINS Laboratories menés dans les années 1965-1070, que la parole n'est pas constituée d'une succession de segments adjacents indépendants correspondant aux phonèmes successifs des mots prononcés. La suite exacte de phonèmesexiste mais à un niveau abstrait dans la pensée du locuteur et dans celle de l'auditeur. Elle n'existe pas vraiment au niveau parole (c'est à dire signal acoustique). A ce niveau, par effet de coarticulation, les phonèmes successifs sont souvent plus ou moins modifiés entre eux (et ce d'autant plus que l'élocution est naturelle et rapide). Cette coarticulation porte sur les traits. Par exemple :



Ce codage, qui n'est pas obligatoire, dépend de 2 contraintes :

- le locuteur cherche à se fatiguer le moins possible,
- tout en émettant le débit d'information le plus grand possible compréhensible par l'auditeur et de ce fait, la coarticulation peut être, pour l'essentiel, décrite par un certain nombre de règles.

Un processus analogue existe pour l'écriture manuscrite, à la différence près que dans ce cas, il existe en plus une version quasi idéale du signal : l'écriture imprimée.

On peut donc dire qu'au niveau acoustique, seuls les traits ont une existence personnalisée : les phonèmes sont très souvent modifiées par les traits des phonèmes voisins et parfois traduits seulement par une modification d'un des traits des phonèmes voisins. Il est donc impossible, à partir du signal acoustique seulement, de reconstituer plus que la suite des traits exprimés. Certains de ces traits permettent de déterminer une suite de phonèmes mais l'utilisation (c'est à dire, la classification exacte) de ces phonèmes n'est possible qu'en faisant intervenir en plus les mots (idéaux) du lexique.

Le système de reconnaissance réalisé à THOMSON-C.S.F.-DASM tient compte de ce phénomène de la façon suivante :

- Une localisation de phonèmes (plutôt qu'une segmentation), est faite au niveau classe de phonèmes (voyelles, semi-consonnes, fricatives, plosives) à partir des traits définissant chacune de ces classes. Intervention du rythme syllabique pour les voyelles.

- Pas de classification définitive des phonèmes localisés, mais pour chaque classe de phonèmes, estimation de quelques paramètres relatifs aux traits permettant d'utiliser ces phonèmes. Par exemple, pour une voyelle, les traits suivants sont utilisés : note de confiance - durée - énergie - degré de nasalisation - description de F1 et F2 - valeur du fondamental.

La figure 1 donne un exemple pour la phrase /5, 2, 5/ de sortie de la chaîne de phonèmes avec pour chaque phonème, le nom de sa classe générale (V pour Voyelle, P pour plosive, sourde ou sonore, N pour plosive nasale, Ø pour fricative, R pour /R/, Z pour silence) et pour chaque classe, un certain nombre de paramètres.

RECONNAISSANCE DE PAROLE CONTINUE MULTILOCUTEUR

SPEAKER INDEPENDANT CONTINUOUS SPEECH RECOGNITION

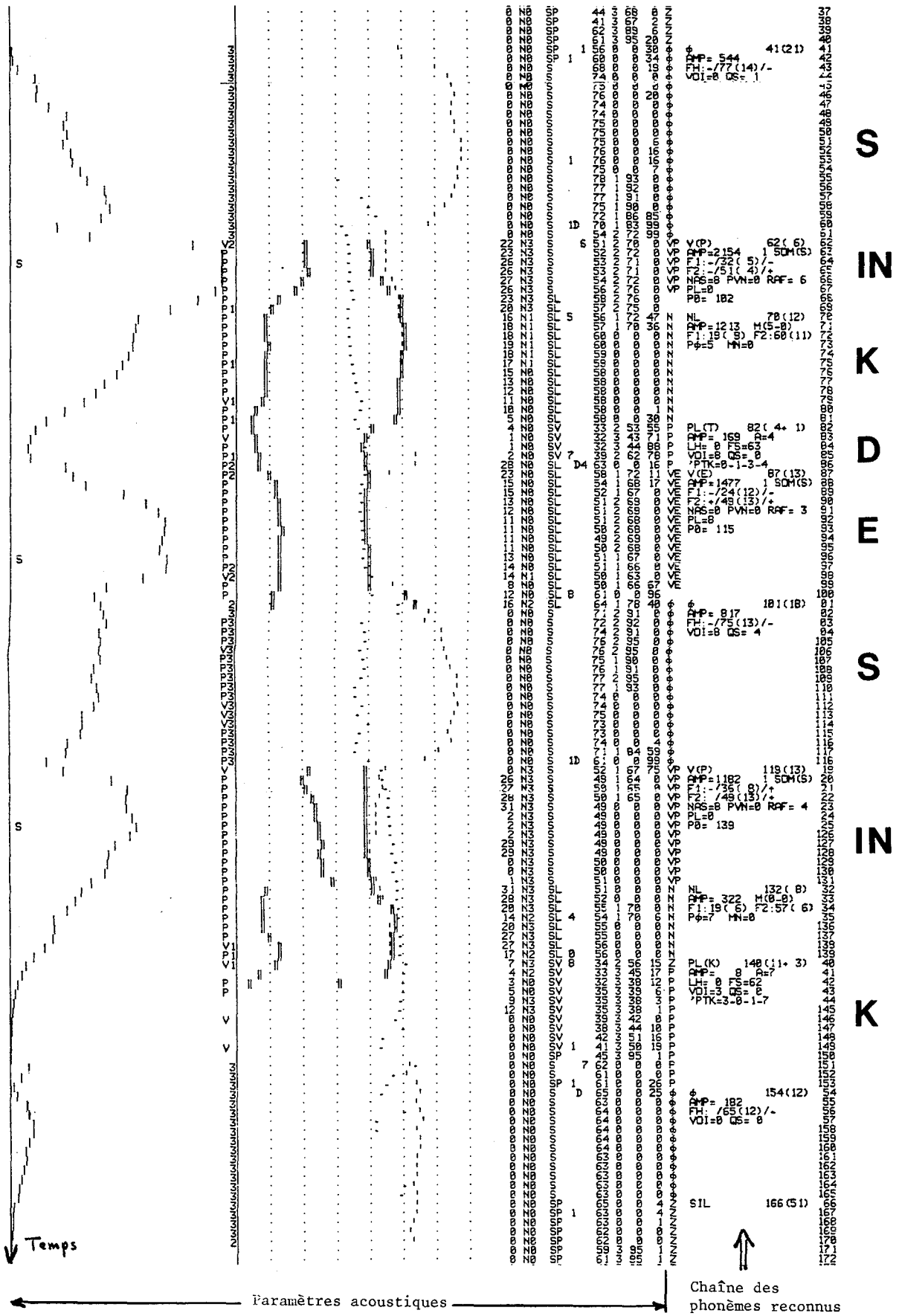


FIGURE 1 : Exemple de chaîne phonémique

5 - COMPARAISON AUX MOTS VOCABULAIRE DE LA CHAINE PHONEMIQUE

Le fait que les phonèmes reconnus soient décri par un ensemble de traits indépendants plutôt que par un label unique complique l'opération de comparaison avec les mots vocabulaire. Les mots vocabulaire sont définis par une chaîne de phonèmes avec indication du degré d'accentuation normal de chaque phonème. Chaque comparaison d'un phonème reconnu avec un phonème vocabulaire donne lieu à une note obtenue en faisant intervenir :

- Des règles relatives à la définition des phonèmes de la langue. Par exemple, un /I/ est une voyelle non nasalisée avec F1 $\in$ [telle zone] et F2 $\in$ [telle autre zone].

Ces règles agissent indépendamment sur chaque trait.

- Des règles de coarticulation qui, en fonction des phonèmes vocabulaire voisins, spécifient des "déformations" possibles des traits (et parfois de la classe) du phonème prononcé.

La technique de la programmation dynamique est utilisée pour adapter au mieux la chaîne de phonèmes reconnus à la chaîne de phonèmes du mot vocabulaire. La note globale mot sert à décider du mot prononcé.

Ce système a été essayé pour faire de la reconnaissance de mots enchaînés (chiffres 0 à 9, alphabet et 30 mots. Exemple de phrase : /Lancer piste Echo 9/). Pour 23 locuteurs ayant prononcé au total 100 phrases (soit 355 mots), 92 % des phrases et 97 % des mots ont été correctement reconnus.

[1] W.A. LEA : "What's wrong with recognition technology" EDD, sept. 82 p. 72.

[2] G. MERCIER : "Evaluation des indices acoustiques utilisés dans l'analyseur phonétique du système KEAL" 9ème JEP du GALF - LANNION 1978 p. 321-342

[3] D.H. KLATT : "Review of the ARPA Speech Understanding Project" JASA vol 62 n° 6 - Dec. 1977 p. 1345-1366.

[4] F. ZURCHER : "Le Vocodeur à canaux : une nouvelle jeunesse ?" CNET Recherche Acoustique vol VI 1979/1980 p. 23 à 40

[5] M. R. SCHROEDER : "Models of Hearing" - Proc. IEEE vol 63 n° 6 sept 1975 p. 1332 à 1350.

[6] P. ALINAT : "Etude des phonèmes de la langue française au moyen d'une cochlée artificielle. Application à la reconnaissance de la parole". Revue technique THOMSON-CSF vol. 7 n° 1 Mars 75

[7] P. ALINAT : Etude du trait permettant de distinguer entre les 3 classes de consonnes explosives PB, TD, KG". 9ème JEP du GALF - LANNION 1978.

[8] K.N. STEVENS, B.E. BLUMSTEIN : "Invarient cues for place of articulation in stop consonants" JASA 64(5) Nov 78, p. 1358-1368.

[9] C.L. SEARLE, J.Z. JACOBSON, S.G. RAYMENT : "Stop consonant discrimination based on human audition" JASA 65(3) Mars 1979 - p. 799-809.

[10] S.E. BLUMSTEIN, K.N. STEVENS : "Acoustic invariance in speech production : Evidence from measurements of the spectral characteristics of stop consonants" JASA 66(4) Oct 79 pp. 1001-1017

[11] A.M. LIBERMAN : "The Grammars of Speech and Language" Cognitive Psychology 1 P. 301-323 1970.

[12] F.S. COOPER :"Acoustics in human communication : Evolving ideas about the nature of speech" - JASA 68(1) - July 1980 pp. 18 à 21

[13] B.H. REPP : "On levels of description in speech research" - JASA 69 (5) May 1981 p. 1462-1464.

[14] A.M. LIBERMAN, F.S. COOPER, D.P. SHANKWEILER, M. Studeent KENNEDY : "Perception of the speeds code" - Psychological review vol. 74 n° 6, Nov. 1967, p. 431 à 461.