Description d'une unité de réponse vocale de données numériques décimales

Maurice Ouaknine, Bernard Teston

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

DESCRIPTION D'UNE UNITE DE REPONSE VOCALE DE DONNEES NUMERIQUES DECIMALES

OUAKNINE Maurice et TESTON Bernard

Laboratoire de PsychoPhysiologie
Université de Provence

Institut de Phonétique
Université de Provence

RESUME

Le système que nous décrivons, est une unité de réponse vocale exclusivement réalisée pour synthétiser des chiffres et des nombres de 1 à 999. Son utilisation est réservée à des applications multimétriques. Elle est connectée à tout appareil disposant des données numérisées en décimal codé binaire. Elle est constitué d'un décodeur lexical, et d'une mémoire de 23 segments de longueur variable qui permettent, présentés dans un ordre adéquat, une synthèse par concaténation. La définition est de 4 bits, le nombre d'échantillons de 3.300 par seconde. On peut envisager très simplement une extension jusqu'aux milliers ainsi qu'une mémoire additionnelle pour stocker quelques unités de mesure.

DESCRIPTION OF A VOCAL RESPONSE UNIT FOR DECIMAL DATA

(1) OUAKNINE Maurice et (2) TESTON Bernard

SUMMARY

The system we describe is a vocal response unit designed sedely for the synthesis of figures and number from 1 to 999, which has been conceived with a view to multemetric application.

It can be connected to any apparatus supplying numerised data in the form of binary coded decimal. It comprises a lexical decoder and a memory of 23 segments of variable length which after appropriate ordering allows synthesis by concatenation.

The definition is of 4 bits and the sampling rate 3.300 per second. The system can easily be extended to bigher numbers and on additionnal memory could used to stock the units of measurement.

(1) Laboratoire de Psychophysiologie
Université de Provence.
(2) Institut de Phonétique
Université de Provence.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

DESCRIPTION D'UNE UNITE DE REPONSE VOCALE DE DONNEES NUMERIQUES DECIMALES.

(1) OUAKNINE Maurice et (2) TESTON Bernard

INTRODUCTION

L'unité de réponse vocale que nous nous proposons de décrire a été développée dans un but bien précis ; transmettre vocalement la mesure d'une grandeur physique par l'intermédiaire d'un multimètre ou tout autre instrument disposant d'une sortie en système décimal codé binaire.

L'application originale d'un tel synthétiseur était, de donner à un sujet, l'information de ses scores au cours d'expériences de psychologie sur la vision, cette information ne devant pas être lue sous peine de perturber les tests. Au début, un opérateur donnait vocalement le résultat des scores aux sujets par l'intermédiaire d'un interphone. Mais, les erreurs de lecture, le retard entraîné par cet intermédiaire humain (tout le reste de la manipulation est automatisé) et les perturbations du sujet au plan de sa concentration, nous ont fait envisager très vite, l'utilisation d'une unité de réponse vocale capable de transmettre toutes les valeurs numériques comprises entre 1 et 999, nécessaires à notre manipulation expérimentale.

La limitation du lexique, ainsi que quelques réalisations industrielles (Master Specialities Co Model 1.700), nous ont fait choisir immédiatement, comme technique de synthèse, la concaténation directe de mots stockés au préalable dans des mémoires numériques.

II - PRINCIPE DE SYNTHESE

La concaténation de segments ou préalablement enregistrés, et découpés en autant d'unité, qu'il apparaissait nécessaire n'a jamais donné de bons résultats pour synthétiser de la parole (CHAFCOULOFF 1976). Pour la langue anglaise, HARRIS (1953) puis WANG et PETERSON (1958) ont avancé le plus dans cette direction, bloquée dès le départ par des problèmes de transitions entre les segments. Cette technique a été complètement abandonnée sauf pour certaines expériences bien particulières (AUTESSERRE et DI CRISTO 1971). Elle a été remplacée avantageusement par les techniques de synthèse par règle. A propos du Français, nous ne connaissons pas de travaux de ce genre dans le passé. Cependant, une tentative actuelle semble se développer sans que nous en ayons connaissance sous forme de publications, mais cela ne tardera-t-il pas malgré le peu d'avenir du système et la médiocrité des résultats.

Si l'on restreint le vocabulaire à concatener à des chiffres et des nombres, on s'aperçoit que l'on n'a pas de problèmes de transition particuliers ; ni de

(1) Laboratoire de PsychoPhysiologie
Université de Provence.
(2) Institut de Phonétique
Université de Provence.

liaison, ni prosodique. Pour compter de 1 à 999 on a besoin que de 23 segments de concaténation. Cette méthode est donc bien adaptée pour résoudre notre problème.

III - DECODEUR LEXICAL

Les 23 segments que nous devons utiliser pour compter de 1 à 999 sont : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 20, 30, 40, 50, 60, 100 et la liaison particulière ET (Tableau N° 1).
Un décodeur lexical particulier est nécessaire pour une bonne numérotation car le système lexical du français des dizaines est particulièrement exceptionnel (dans le sens où il existe de nombreuses exceptions à la règle générale). Sur ce plan, l'énumération en langue anglaise ou allemande est plus simple. Il nous faut réaliser de nombreux masquage selon les ordres de sortie des groupes de chiffres. (Tableau 2)
Le décodage du nombre (digit) en rapport au nombre effectivement prononcé est parfois complexe (Exemple : 97 → 4, 20, 10, 7) et le décodeur lexical que nous avons voulu le plus simple possible nous a occasionné quelques difficultés. Plutôt qu'une explication confuse, nous reportons le lecteur au tableau 2 pour en saisir le principe de fonctionnement, ainsi qu'à la figure 3.

Les informations numériques, images des grandeurs auparavant mesurées, représentée en décimal codé binaire, sont envoyées en unités, dizaines et centaines, dans sept registres d'adresse sur 5 bits et un registre à décalage de 1 bit pour l'inhibition des chiffres qui ne doivent pas être émis vocalement (Figure 4). Le code binaire sur 5 bits des différents segments ou groupe de chiffres est donné au Tableau 5.

IV - SYNTHETISEUR

Il est essentiellement constitué par une mémoire de 23 mots d'un nombre de bits variables, plus ou moins important selon la durée du segment.
Cette mémoire morte est constituée par 23 boitiers de 4 x 1.024 bits.

Les adresses des mots ainsi que le balayage des bits successifs constituant ces mots sont réalisés au moyen de compteurs - décompteurs synchrones, comparateurs et décodeurs.

La sortie des informations se fait sur 4 bits qui attaquent un convertisseur numérique - analogique suivi d'un filtre passe bande dissymétrique (150 Hertz - 36 dB/octave - 2,5 K Hertz - 96 dB/octave). L'horloge de lecture bat à la fréquence de 3.300 Hertz (Figure 6).

V - PREPARATION DU CORPUS DE SYNTHESE

La durée des différents chiffres et nombres étant très variable (Tableau 7) nous avons enregistré en chambre sourde, 10 locuteurs choisis dans notre entourage. Les 23 séquences ont été prononcées naturellement sans accentuations particulières, recto tono et bien séparées par un important silence (2 à 3 secondes). Ceci fait, nous avons réalisé une moyenne sur les différents segments, et nous avons choisi le locuteur qui se rapprochait le plus de cette moyenne mais aussi dont la voix restait esthétique et intelligible malgré les traitements que nous lui faisions subir.

Le cadrage à la longueur exacte des segments a été effectué au moyen d'un adressage numérique sur une RAM après conversion sur 8 bits du signal analogique. La diminution n'a été que de quelques bits afin d'assurer un bon cadrage en fonction du nombre de bits disponible dans la mémoire (Tableau N° 8). La translation cadrée en durée sur les PROM a été réalisée manuellement sur 4 bits. L'opération

s'est avérée longue et fastidieuse, mais nous ne pouvions pas utiliser notre calculateur pour ce premier essai. La fréquence d'échantillonnage a été choisie à la valeur de 3.300 Hertz après avoir tenu compte de la durée totale des segments et de notre capacité de mémoire maximale.

Avant de mémoriser définitivement les segments sur les PROM, nous avons testé différents procédés dans le but d'améliorer la qualité de la voix restituée dans les conditions de compression d'information précisées précédemment.

Tout d'abord, nous avons essayé de doubler la fréquence d'échantillonnage pour augmenter la largeur de bande du spectre des consonnes constructives surtout dans 6 (six) et 10 (dix). Les essais furent concluants et l'amélioration de la distinction de ces deux chiffres très sensibles. Cependant, le fait de devoir, outre la fréquence d'échantillonnage, changer la valeur de la fréquence de coupure du filtre passe bas de sortie, ainsi que le doublement de la capacité mémoire de ces deux chiffres nous ont dissuadé d'employer cette solution. Toujours dans le même sens, nous avons également testé un système d'échantillonnage continuement variable en fonction de la fréquence des signaux à échantillonner. Ce système est très efficace pour optimiser l'encombrement des mémoires, mais il est compliqué à mettre en oeuvre.

Pour augmenter la dynamique de l'amplitude des signaux, nous avons essayé de comprimer puis d'expanser le signal au moyen de circuits appropriés analogiques, mais nous ne les avons pas retenus car ils compliquaient par trop le système, malgré une nette amélioration de la dynamique de synthèse.

VI - RESULTATS

Les résultats que nous obtenons, avec les durées du Tableau 8, 4 bits de définition du signal et 3.300 Hertz de fréquence d'échantillonnage sont satisfaisants. Le taux d'erreur de compréhension des valeurs numériques ainsi transmises aux sujets est inférieur à 3 %. Dans la version actuelle, nous avons supprimé le ET de liaison (Exemple : 20 et 1, 30 et 1 etc...). Cette simplification importante ne semble pas perturber la compréhension. Ceci malgré la présence dans les segments de nombreuses erreurs de programmation de la PROM, dues à la manipulation manuelle des données. Le décodeur lexical fonctionne parfaitement et semble ne pas pouvoir être plus simple.

La qualité de la voix ainsi restituée peut être avantageusement améliorée en utilisant pour les conversions AN et NA, des convertisseurs à compression-expension de dynamique (P.M.I. type DAC 76 par exemple) qui viennent d'apparaître sur le marché. On peut envisager avec ces systèmes de 4 bits effectifs, d'obtenir une dynamique de 6 bits. On peut également augmenter la largeur du spectre jusqu'à la bande passante téléphonique, en doublant la capacité des mémoires, car ces dernières deviennent de plus en plus compactes et coûtent de moins en moins cher.

VII - CONCLUSION

Le système de réponse vocale que nous venons de décrire peut être très facilement étendu pour devenir un système complet pour multimètre numérique.

Le décodeur lexical comprend déjà les milliers, ce qui permet de compter jusqu'à 999.999 avec un mot de mémoire supplémentaire (1.000). Pour des applications multimétriques, une mémoire spéciale pour le stockage des unités peut être réalisée très simplement (Hertz - Volts - Décibels - Millibars etc...). De tels systèmes de réponse vocale peuvent être utilisés dans de nombreuses occasions. Dans certaines conditions de travail difficile, où toute l'attention d'un sujet

est prise par une tache principale nécessitant la connaissance de nombreux paramètres (pilotage d'hélicoptères - interventions chirurgicales, maintenance technique dans de mauvaises conditions d'accessibilité etc... etc...). On peut également envisager ainsi, la scrutation de données centralisées au moyen d'une simple ligne téléphonique à grande distance.

Dès maintenant, un système de 4 bits de définition avec compression expension de dynamique, comptant de 1 à 999.999 et d'une bande passante de 100 à 3.300 Hertz peut être réalisé avec moins d'une vingtaine de boîtiers de C.I. logiques.

On peut envisager même un circuit complexe contenant sur la même puce tout le synthétiseur de grandeur numérique décimale. Le constructeur pourrait à la demande proposer trois types de voix ; homme, femme, et type "Aéroport" pour horloge parlante publique par exemple. Le marché d'un tel composant existe dès maintenant. Son coût de production devra pourtant être comparé à des systèmes de synthèse à prédiction linéaire qui viennent de sortir sur le marché nord américain (WIGGINS, R. 1978). Si l'on se place strictement dans notre application, il nous semble que la synthèse par concaténation a des chances de se développer.

-

10^1 \ 10^0	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
1	10	11	12	13	14	15	16	10 → 7	10 → 8	10 → 9
2	20	20 et 1	20 → 2	20 → 3	20 → 4	20 → 5	20 → 6	20 → 7	20 → 8	20 → 9
3	30	30 et 1	30 → 2	30 → 3	30 → 4	30 → 5	30 → 6	30 → 7	30 → 8	30 → 9
4	40	40 et 1	40 → 2	40 → 3	40 → 4	40 → 5	40 → 6	40 → 7	40 → 8	40 → 9
5	50	50 et 1	50 → 2	50 → 3	50 → 4	50 → 5	50 → 6	50 → 7	50 → 8	50 → 9
6	60	60 et 1	60 → 2	60 → 3	60 → 4	60 → 5	60 → 6	60 → 7	60 → 8	60 → 9
7	60 → 10	60 → 11	60 → 12	60 → 13	60 → 14	60 → 15	60 → 16	60 → 17	60 → 18	60 → 19
8	4 20	4 → 20 → 1	4 → 20 → 2	4 → 20 → 3	4 → 20 → 4	4 → 20 → 5	4 → 20 → 6	4 → 20 → 7	4 → 20 → 8	4 → 20 → 9
9	4 20 10	4 → 20 → 11	4 → 20 → 12	4 → 20 → 13	4 → 20 → 14	4 → 20 → 15	4 → 20 → 16	4 → 20 →10 → 7	4 → 20 → 10 → 8	4 → 20 → 10 → 9

- Composition des nombres de 1 à 99.

Tableau N° 1

- Number composition (1 to 99).

Ordre d'apparition des groupes de chiffres		Commentaire particulier de masquage
A) 1,2,3,4,5,6,7,8,9	10^2	1 masqué
B) 100		0 masqué
C) 4	10^1	Masqué sauf pour 8 + 9
D) 20,30,40,50,60		20 sort pour : 8 + 9 (2 + 8 + 9)
		60 pour 7 (6 + 7)
E) ET → $(n)_{10}1$ $(1)_{10°}$		Masqué pour $(0)_{10}1$ et $(1)_{10}1$ $(1)_{10°}$
F) 10,11,12,13,14,15,16		Sort pour (1 + 7 + 9) x $(0 + 1 + 2 + 3 + 4 + 5 + 6)_{10°}$ $(1 + 7 + 9)_{10}1$ x $\overline{(7 + 8 + 9)_{10°}}$
G) 1,2,3,4,5,6,7,8,9		Sort lorsque F est masqué

- Ordre de sortie des différents groupes de chiffres avec masquage éventuel.

Tableau N° 2

- Output sequence of different number groups with eventual mask.

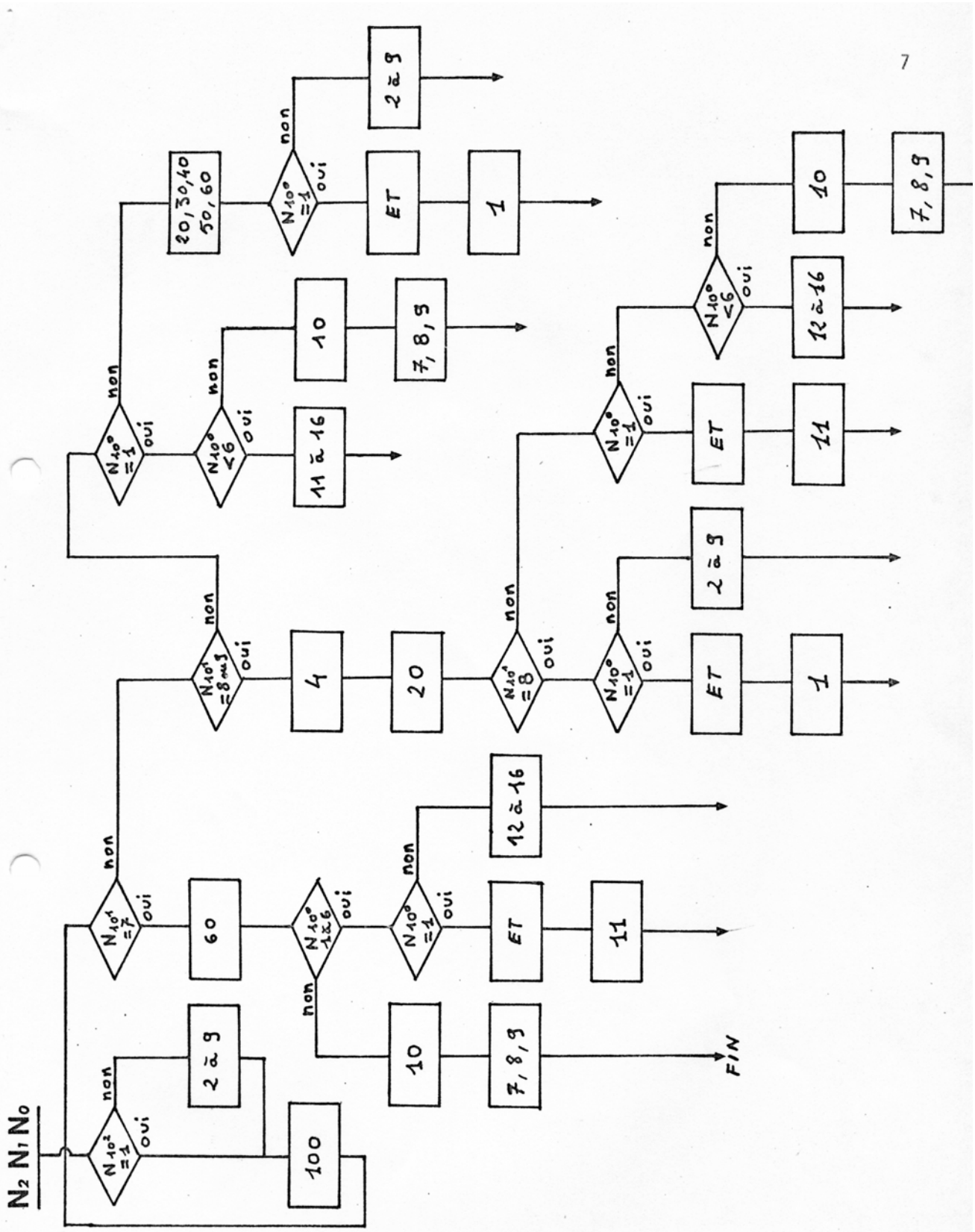

Fig. 3 - Ordinagramme du décodeur lexical de 1 à 999.
- Flowchart of the lexical decoder (1 to 999).

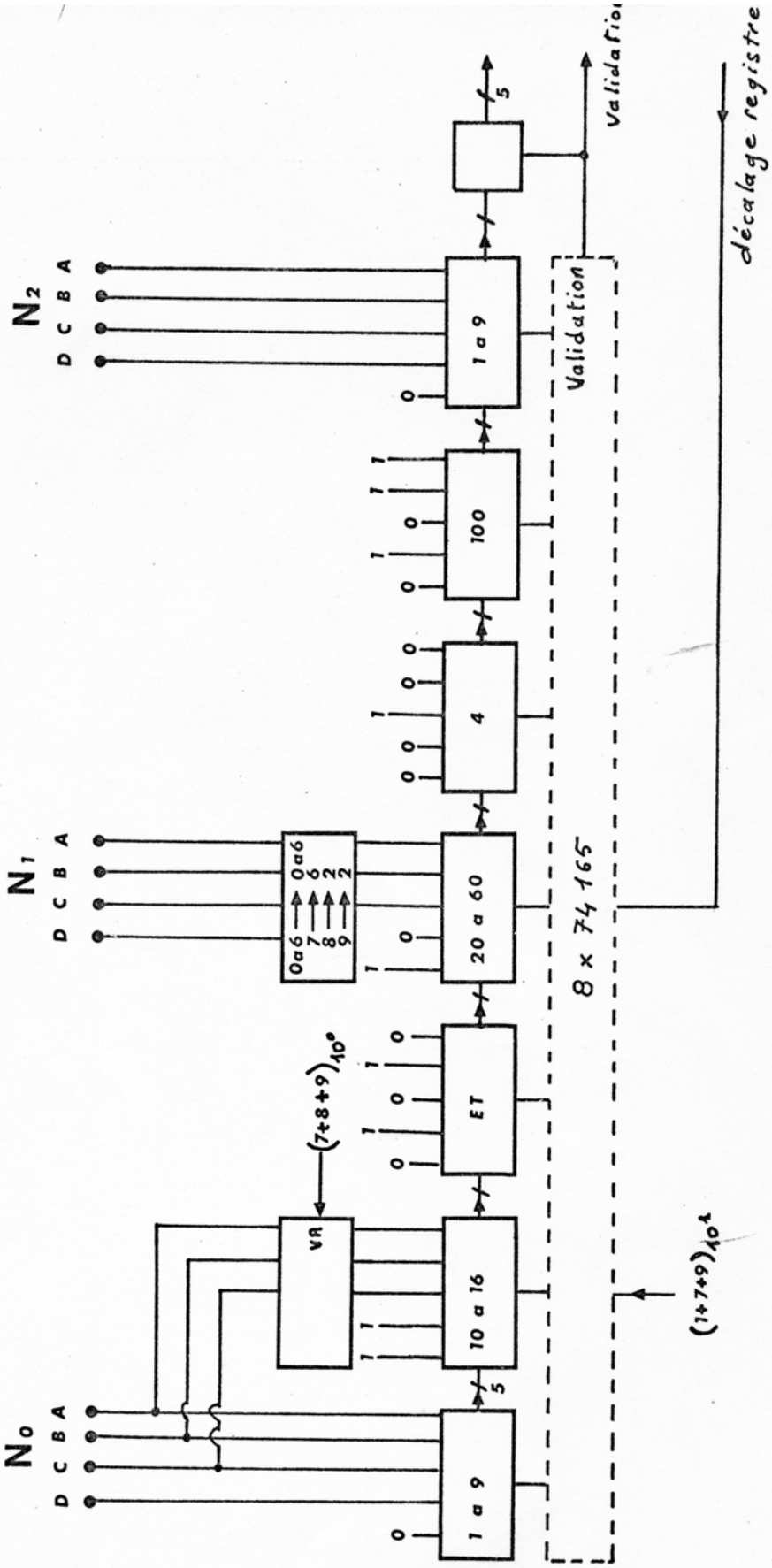

Fig. 4 - Schéma de principe du décodeur lexical.

- General Outline of the lexical decoder.

Groupe de chiffre	Chiffre ou Nombre (23)	Code Binaire	
A	1	0 0 0 0 1	
	2	0 0 0 1 0	
	3	0 0 0 1 1	
	4	0 0 1 0 0	
	5	0 0 1 0 1	
	6	0 0 1 1 0	
	7	0 0 1 1 1	
	8	0 1 0 0 0	
	9	0 1 0 0 1	
B	100	0 1 0 1 1	
C	4	0 0 1 0 0	
D	20	1 0 0 1 0	$(2 + 8 + 9)_{10}{}^{1}$
	30	1 0 0 1 1	
	40	1 0 1 0 0	
	50	1 0 1 0 1	
	60	1 0 1 1 0	$(6 + 7)_{10}{}^{1}$
E	ET	0 1 0 1 0	

- Code binaire des différents segments.

Tableau N° 5

- Binary code of the different segments.

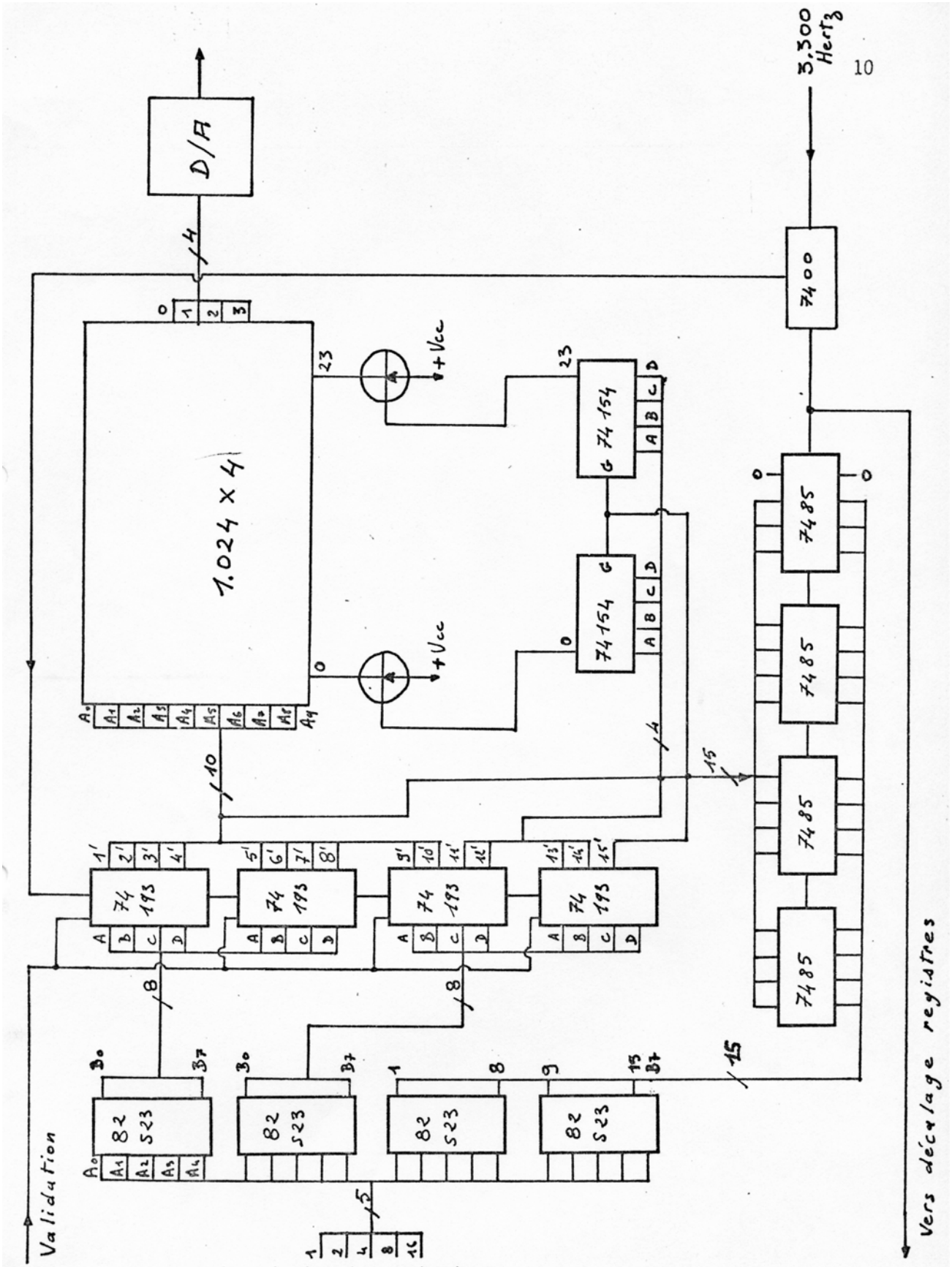

Fig. 6 - Schéma de principe du synthétiseur.

- General Outline of the synthesizer.

N° du Locuteur	Esthétique	Intelligibilité	Durée Moyenne	Durée du segment 5 (cinq) en milli-secondes
1				450
2				500
3		x	x	350
4			x	375
5				450
6				400
7	x	x	x	375
8				385
9			x	375
10				410

- Durée du chiffre cinq prononcé par 10 locuteurs différents.

Tableau N° 7

- Duration of the number five, pronounced by 10 different speakers.

SEGMENT (chiffre ou nombre)	DUREE CHOISIE en millisecondes	NOMBRE DE BITS dans la Mémoire
1	231	765
2	189	624
3	330	1.023
4	310	1.023
5	370	1.217
6	280	924
7	250	831
8	314	1.023
9	310	1.023
10	239	791
11	300	975
12	314	1.023
13	350	1.151
14	566	1.871
15	355	1.167
16	310	1.023
20	250	835
30	310	1.023
40	430	1.415
50	540	1.783
60	465	1.535
100	250	831
1000	465	1.535

- Durée moyenne choisie pour la durée des segments et leur taille mémoire respective.

Tableau N° 8

- Mean duration elected for the segment duration and the memory wide.

REFERENCES

AUTESSERRE, D. et DI CRISTO, A., 1972, "Recherche sur l'intonation du Français : Traits significatifs et non significatifs", Proceedings of the VIIth International Congress of Phonetic Sciences, Mouton, The Hague, 1972, pp. 842-859.

CHAFCOULOFF, M., 1976, Vingt cinq années de recherches en synthèse de la parole, Editions du C.N.R.S., Paris, p. 283, 1976.

HARRIS, C.M., 1953, "A Study of the building blacks in speech", J.A.S.A., 25, 5, pp. 962-969.

HNATEK, E.R., 1976, A user's hondboak of semiconductor memories. J. Wiley, Londres, p. 688, 1976.

HNATEK, E.R., 1977, A user's hondboak of D/A and A/D converters. J. Wiley, Londres, p. 472, 1977.

LEE, S.C., 1976, Digital circuits and logic design Prenctice Hall, New York, p. 594, 1976

ROSS, E.J., 1977, Modern digital communications Mc Grow Hill, New-York, p. 308, 1977

WANG, W., PETERSON, G.E., 1958 "Segment Inventory for Speech Synthesis". J.A.S.A., 30, 8, pp. 743-746, 1958.

WIGGINS, R., BRANTINGHAM, L., 1978, "Three chip system synthesizes human speech" Electronics 51,18, August, pp. 109-116.