Chapitre 4

Vocabulaire et phonétique

Nous avons étudié, dans les chapitres précédents, les principes généraux de la synthèse de parole, leur mise en œuvre dans le MEA 8000, l'interface avec le microprocesseur et le logiciel de base permettant la prononciation d'une expression existant déjà en mémoire. Nous allons maintenant voir de quelle manière ce vocabulaire peut être constitué en fonction de l'objectif à atteindre, ainsi que les avantages et inconvénients respectifs des différentes possibilités.

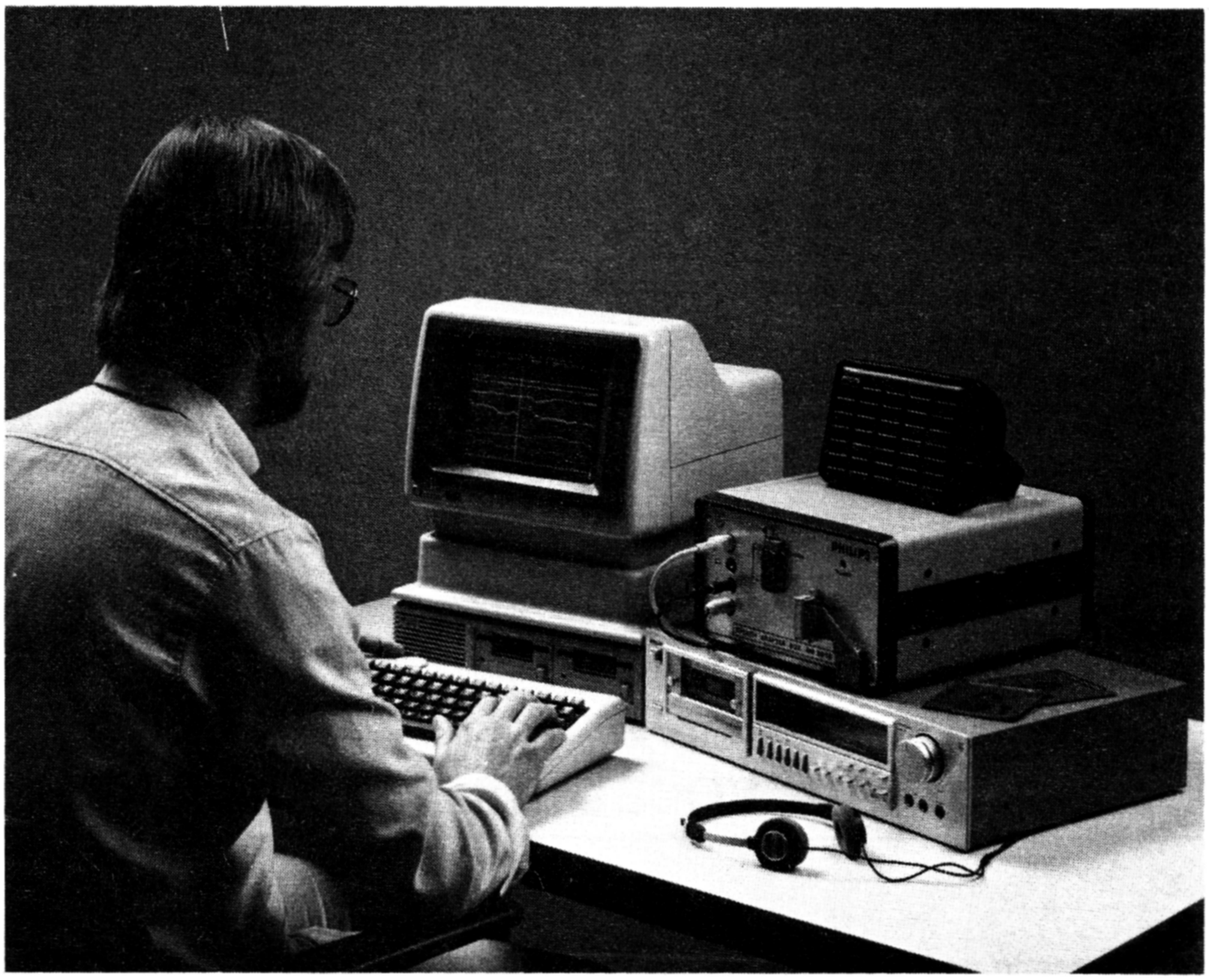
Création du vocabulaire codé

Pour créer le vocabulaire codé au format que nous avons vu au chapitre précédent, il faut disposer d'un système d'analyse et de codage de la parole tel que celui proposé par RTC, visible sur la photographie. Ce système se compose de deux éléments principaux:

— Un ordinateur 16 bits muni d'une interface IEEE 488 (= CEI 625, HPIB, GPIB) sur lequel un logiciel spécifique sera exécuté. Aujourd'hui, ce logiciel existe pour deux types de machines (IBM PC et HP 9816S).

— Une boîte d'adaptation (SAB) connectée à l'ordinateur par le BUS IEEE 488 et incorporant 5 blocs fonctionnels (interface d'entrée-sortie et conversion A/N et N/A, interface IEEE 488, synthèse de parole, programmation d'EPROM, alimentation).

Cet ensemble muni de son logiciel permet, à partir d'un enregistrement magnétique, la réalisation de toutes les étapes nécessaires à la création du vocabulaire codé en EPROM; la phase d'édition (correction /compression) qui est décisive pour la qualité du résultat, est grandement facilitée par la représentation graphique de tous les paramètres. Selon le mode de compression choisi (automatique ou manuel), les débits d'information seront compris entre 1 000 et 2 000 bits par seconde de parole. La partie spécifique (SAB et logiciel) est commercialisée par le fabricant sous la référence OM 8010.



Système de codage RTC OM 8010, ici en fonctionnement sur le micro-ordinateur HP 9816S

Quel vocabulaire ?
Sous quelle forme ?

Le vocabulaire peut être stocké sous l'une des formes suivantes:

— Phrases complètes préenregistrées et codées spécialement pour une application déterminée. Cette méthode est celle qui donne, du point de vue de la qualité acoustique globale, les meilleurs résultats car elle conserve les caractéristiques prosodiques (intonation et rythme) du message d'origine. Elle a pour inconvénient principal de nécessiter un codage spécifique relativement coûteux générant un vocabulaire figé, limité à l'application prévue. De ce fait, cette méthode est en général réservée aux applications de moyenne ou grande série, telles que l'automobile par exemple.

— Mots ou éléments de phrase isolés permettant par combinaison de constituer un certain nombre de messages. Une application assez courante de cette méthode consiste à utiliser une partie de phrase fixe avec une portion variable, constituée de chiffres par exemple. Cette méthode, si elle est un peu plus souple que la précédente, donne d'un peu moins bons résultats; en effet, d'une part un mot n'est pas prononcé de la même manière selon sa position dans la phrase, et d'autre part les liaisons existant entre les mots en français, notamment dans les chiffres composés, ne facilitent pas les choses... Le vocabulaire fourni en annexe permet néanmoins d'intéressantes applications de ce type (répondeur, horloge parlante, etc.) dont certains exemples seront décrits.

— Eléments phonétiques prédéfinis (phonèmes, diphones): c'est cette dernière méthode qui retiendra le plus notre attention car c'est la seule qui permette l'accès à un vocabulaire illimité (dans une langue déterminée) sans nécessiter de codage préalable (synthèse à partir du texte). Elle implique de disposer d'un "dictionnaire" d'éléments phonétiques et des règles pour s'en servir, qui détermineront la qualité des résultats; quelques notions de phonétique sont nécessaires pour l'utiliser.

Un peu de phonétique

Toute langue parlée n'utilise qu'un nombre relativement limité de sons élémentaires qui permettent par leur concaténation de composer tous les messages imaginables dans cette langue. On appelle "phonèmes" ces sons élémentaires dont le nombre peut varier d'une trentaine à une soixantaine selon la langue ou le dialecte considérés. La liste de ces phonèmes varie notablement d'une langue à l'autre, même pour des langues d'origine commune, latine ou saxonne par exemple. C'est ce qui explique les résultats très médiocres obtenus lorsque l'on essaie de synthétiser un texte à partir d'éléments phonétiques développés pour une autre langue. En ce qui concerne le français, on compte généralement 37 phonèmes, dont certains sont proches (par exemple les sons "in", "ein", "ain", "un", "eun"... que la pratique quotidienne tend à confondre).

L'utilisation de phonèmes offre l'avantage de ne nécessiter qu'un dictionnaire restreint, donc peu encombrant en mémoire, et d'être simple d'utilisation si on se contente d'une entrée phonétique des expressions à vocaliser. C'est pourquoi les principales applications décrites dans cet ouvrage, dont la destination première est l'initiation en vue d'utilisations individuelles, sont basées sur cette méthode.

Il faut cependant reconnaître que la qualité obtenue avec la synthèse par phonèmes est, par son principe, limitée. En effet, le phonème est une entité théorique qui n'existe pas dans le langage parlé réel, en raison du phénomène connu sous le nom de "coarticulation" : chaque son émis est influencé par son prédecesseur et son successeur, du fait même des caractéristiques mécaniques de l'appareil vocal. Ces zones de transition entre un son et un autre sont très importantes, et on est conduit pour les respecter à définir un autre élément phonétique: le diphone ou diphonème.

Celui-ci s'étend du milieu de la partie "stable" d'un son au milieu de la partie stable de son successeur. On voit donc, si l'on considère les 37 sons élémentaires du français, qu'il y a 37 x 37 = 1369 combinaisons, donc diphones, différents. En pratique, certains sont très voisins ou inusités, et l'on considère que 1200 diphones permettent d'obtenir de très bons résultats. On imagine toutefois sans difficulté que l'encombrement d'un dictionnaire de diphones sera environ 30 fois supérieur au dictionnaire de phonèmes correspondant, et que le logiciel d'utilisation sera plus complexe.

En France, le CNET (Centre National d'Etudes des Télécommunications) de Lannion a développé des dictionnaires de diphones et des logiciels de synthèse à partir du texte écrit, utilisables sur un synthétiseur spécifique. Outre les résultats qualitatifs dus à l'utilisation de diphones et à la génération d'une prosodie correcte, le logiciel de conversion orthographique/phonétique utilisé permet une entrée du texte sous forme habituelle ; il lève en effet la plupart des ambiguïtés de prononciation telles que celle de la célèbre phrase "les poules couvent au couvent". Ces résultats sont le fruit de nombreuses années de travaux d'équipes pluridisciplinaires et ne sont donc pas du domaine public.

Une adaptation au MEA 8000 sur un micro-ordinateur APPLE II en a été faite par la société MATRA; elle est constituée d'une carte et d'un logiciel spécifiques commercialisés sous le nom de "PORTE-PAROLE".

Un dictionnaire de diphones, développé par l'IPO d'Eindhoven (Hollande), existe également sur le MEA 8000 pour le néerlandais.

Une autre méthode encore peu développée en France pour la synthèse à partir du texte est la synthèse "par règles" qui, à partir d'un dictionnaire de phonèmes calcule l'évolution (trajectoire) entre les paramètres de deux phonèmes successifs au moyen de modèles mathémathiques ou "règles". Cette méthode a généralement recours aux formants comme paramètres, et a été utilisée avec succès aux U.S.A., en Suède et au Canada.

La prosodie ?
Ce n'est pas si simple...

Ce que l'on appelle la prosodie d'une phrase, c'est en fait essentiellement l'information supplémentaire transmise par un message parlé par rapport au même message écrit, c'est-à-dire l'intonation, l'accent tonique et le rythme de la phrase. Les principaux paramètres influençant la prosodie sont :

— La variation du fondamental (pitch) est le paramètre déterminant l'intonation, et la même phrase peut voir son caractère passer du mode affirmatif au mode interrogatif par sa seule modification. C'est le paramètre le plus important pour la prosodie globale d'une expression.

— Les variations locales de l'amplitude caractérisent l'accent tonique. Celui-ci est toutefois peu marqué en français, où toutes les syllabes sont prononcées avec des intensités assez constantes. Ce paramètre est donc relativement secondaire pour la synthèse.

— La durée des syllabes successives définit le rythme de la phrase, qui pourra être modifié en jouant sur la durée des phonèmes la composant.

Si les paramètres déterminant ces caractéristiques sont relativement bien connus, les lois régissant leurs variations en fonction de la progression de la phrase sont assez complexes, et nous nous bornerons à fournir quelques règles générales permettant d'ajouter un peu de naturel aux messages construits à partir de phonèmes, ou de modifier l'intonation d'un message prédéfini.

Intonation

Le fondamental, qui rappelons-le est la fréquence de vibration des cordes vocales lors des sons voisés, varie entre 70 et 200 Hz pour l'homme adulte et entre 100 et 300 Hz pour la femme adulte dans la conversation courante. Il peut dépasser largement ces limites lors du chant ou de situations particulières. Dans une phrase, la courbe de variation de ce paramètre n'est en général pas uniforme et présente de nombreux changements de pente; nous allons essayer de donner quelques indications permettant d'établir une telle courbe.

• Allure générale de la courbe: on a remarqué qu'en général la courbe d'évolution du fondamental présentait une allure descendante au long de la phrase; ceci semble vérifié dans toutes les langues, et donc probablement dû à des causes physiologiques. Le fait de créer une telle "ligne de déclinaison" donnera donc un certain naturel à la parole synthétique. Autour de cette ligne, on observe des variations de pente locales au niveau des syllabes composant le message.

• Variations de pente locales: on a observé depuis très longtemps que, dans une phrase de structure simple, si celle-ci était sur le mode affirmatif, la pente de la courbe d'intonation de la dernière syllabe, accentuée, était négative; à l'inverse la pente est positive si le mode est interrogatif. Ainsi la phrase "Il fait beau!" aura une pente descendante sur "beau" alors qu'une pente montante sur "beau" évoquera une interrogation:

	Il fait BEAU
Mode affirmatif	\ !
Mode interrogatif	/ ?

De nombreux chercheurs se sont efforcés de déterminer les règles gouvernant les variations de pente locales à l'intérieur de la phrase. Les résultats de ces recherches ont permis d'établir certains liens entre la syntaxe et la prosodie de la phrase; ces règles sont complexes et donc difficilement utilisables par un amateur. Il a cependant pu être établi que les variations de pente sont régies par une opposition entre la pente de la syllabe accentuée terminant la phrase (déterminée par le mode affirmatif ou interrogatif) et celle de la syllabe accentuée du mot lui faisant face.

Par exemple, pour la phrase "Geneviève partira" les syllabes accentuées sont montrées en majuscule, et la pente correspondante est indiquée au dessous pour les deux modes possibles:

	geneVIEve partiRA
Mode affirmatif	/ \ !
Mode interrogatif	\ / ?

Ceci est relativement simple à mettre en œuvre sur des phrases courtes du type "sujet-verbe" telle que celle ci-dessus, où l'opposition se fait entre sujet et verbe, et "sujet-verbe-complément", où l'opposition a lieu entre sujet et complément, comme ci-dessous:

	geneVIEve partira deMAIN
Mode affirmatif	/ \ !
Mode interrogatif	\ / ?

Dans des phrases plus longues, le principe est le même, mais on devra distinguer les oppositions à plusieurs niveaux, ce qui rend l'opération beaucoup plus délicate.

Il n'est pas facile de mettre en évidence ce type de structure dans une phrase un peu compliquée, non plus que de découvrir les syllabes accentuées dans la phrase. Dans beaucoup de cas, on devra procéder par essais et erreurs. On aura toujours la possibilité d'imprimer une ligne de déclinaison et de fixer la pente de la dernière syllabe en fonction du mode (affirmatif ou interrogatif) recherché, si l'on veut éviter la "voix de robot".

Rythme

La durée relative des syllabes d'une phrase en détermine le rythme qui est un facteur important du naturel du message. Les lois régissant ces durées sont encore mal connues et si l'on désire travailler ce point, il faudra procéder par tâtonnements. On pourra pour cela ajouter ou retrancher des trames aux syllabes (essentiellement à la voyelle la composant), ou modifier la durée de certaines trames.

Avec la version de phonèmes 4.2, codée avec une durée de trame fixe de 16 ms, il sera également possible de modifier la vitesse globale d'élocution en changeant la durée de trames sur l'ensemble de la phrase. On pourra ainsi doubler ou diviser par deux la vitesse en passant toutes les trames à 8 ou 32 ms respectivement; des vitesses intermédiaires pourront être obtenues en ne faisant cette opération qu'une trame sur deux.

Les phonèmes du MEA 8000

Nous avons vu que le français utilisait théoriquement 37 phonèmes pour composer son vocabulaire. Dans la pratique, nous en avons identifié 31 suffisamment différenciés par l'usage courant pour nécessiter un code particulier. Nous avons cependant trouvé intéressant d'ajouter à ces phonèmes 7 sons composés (diphtongues) dont la reproduction par concaténation de 2 phonèmes n'était pas très satisfaisante. D'autre part, deux silences de durée 32 et 64 ms ont été ajoutés pour servir de séparateurs, par exemple comme des signes de ponctuation. La liste de ces 40 éléments avec leur représentation écrite et un exemple d'utilisation pour ceux dont la forme écrite est ambiguë est fournie en annexe p. 258.

Dans les applications sur les diverses machines, nous utiliserons un caractère pour représenter chaque phonème, en utilisant la lettre correspondante de l'alphabet lorsque la correspondance entre lettre et phonème est unique. Pour les sons habituellement représentés par un groupe de lettres, nous utiliserons l'un des symboles spéciaux du clavier. Toutes les machines ne disposant pas exactement du même jeu de caractères, cette liste variera légèrement de l'une à l'autre.

Tous les phonèmes sont codés avec un pitch constant (120 Hz) afin qu'ils puissent s'enchaîner dans un ordre quelconque sans discontinuité ; ceci a pour effet de produire une "voix de robot" à laquelle il sera possible de donner une intonation artificielle par l'utilisation de marqueurs dans l'expression crée.

Afin de permettre une concaténation satisfaisante, les phonèmes ont été normalisés en amplitude (avec une croissance, une partie plate et une décroissance) et en durée (en général 128 ms pour les voyelles). Pour les consonnes plosives (p, t, k, b, d, g) celles-ci sont toujours précédées d'un silence dû à l'occlusion de la bouche qui les précède ; c'est pourquoi un silence de 32 ms est codé au début de ces phonèmes (version 4.2).

Enfin, deux tables de phonèmes différentes ont été codées :
— la première (version 3.3) a été codée avec une durée de trame variable de façon à minimiser son encombrement, et occupe exactement 1 kilo-octet ;
— la seconde (version 4.2) a été codée avec une durée fixe de 16 ms, ce qui augmente son encombrement de près de 50%, mais ajoute d'intéressantes possibilités, comme par exemple la variation de la vitesse d'élocution par modification de la durée de trame ; la trame fixe facilite également la conception d'un programme d'édition graphique. Ces deux tables utilisant les mêmes numéros de code pour leurs phonèmes sont compatibles en logiciel, à l'encombrement près.

Phonèmes, mode d'emploi

Pour utiliser ces phonèmes, deux possibilités (au moins) nous sont offertes :

— fabriquer par concaténation une expression à un emplacement mémoire déterminé, en allant chercher les codes des phonèmes dans la table. Cette expression sera codée selon le format décrit au chapitre précédent (en-tête de 4 octets). Ceci peut être fait par un programme BASIC qui appellera la routine en langage-machine de commande du MEA 8000 pour la vocalisation ;

— vocaliser les phonèmes directement où ils se trouvent dans la table, leur séquence déterminant l'expression prononcée. Ceci implique, pour des raisons de rapidité, un programme entièrement en langage-machine, pouvant être appelé par un programme utilisateur, en BASIC par exemple.

Chacune de ces deux méthodes présente ses avantages propres, et les deux solutions seront proposées.

La première solution permet de modifier certains paramètres pour optimiser la prosodie de la phrase (l'intonation avec le pitch, le rythme en ajoutant ou retranchant des trames), et d'obtenir des effets particuliers (chuchotement par "dévoisement", vitesse d'élocution en modifiant la durée de trame). L'expression ainsi éditée pourra être sauvegardée et rappelée par un programme utilisateur. Un programme mettant en œuvre cette solution est fourni dans les chapitres suivants pour chacune des machines étudiées.

La deuxième solution offre l'avantage d'une prononciation immédiate de l'expression créée, et de ne pas nécessiter d'autre espace mémoire que celui utilisé par le dictionnaire de phonèmes et le programme L.M., quels que soient le nombre et la longueur des expressions désirées. En contre-partie, la voix sera très "robotique", aucune édition n'étant possible ici. Un programme de ce type est donné pour chacun des microprocesseurs étudiés ici (6502, Z-80, 6809).

Enfin un programme en langage-machine pour les MO5, TO7, TO7/70 baptisé "Phonetram" et permettant une édition graphique de tous les paramètres, est abondamment détaillé.

Et le chant ?

Le chant est un cas particulier de parole dans lequel la hauteur n'est pas déterminée par l'intonation du message, mais par la mélodie. Le texte à chanter est une suite de syllabes qui représentent chacune une note dont la hauteur et la durée sont indiquées dans la partition. Par exemple, pour le début de "Au clair de la lune", la séquence est:

Texte	Au	clair	de	la	lu – ne
Notes	do	do	do	ré	mi ré
Durée	Nr	Nr	Nr	Nr	Bl Bl

(Nr = noire. Bl = blanche)

Donc pour la création d'un chant à partir d'un texte composé au moyen de phonèmes, il faudra d'une part adapter la durée des syllabes et d'autre part leur donner la hauteur correspondant à la note à jouer. Avec le MEA 8000, il sera possible de créer une expression correspondant au texte de la chanson au moyen des phonèmes, puis de modifier la hauteur de chacune des syllabes pour la faire correspondre à la note désirée, et enfin de définir la durée de chaque note soit par modification de la durée de trame, soit en ajoutant ou retranchant des trames a la voyelle caractérisant la note.

En pratique, la hauteur ne pourra varier que sur deux octaves environ (80 à 400 Hz avec la voix utilisée pour les phonèmes) pour rester "vraisemblable". La gamme de variation de durée sera dans un rapport 8 en jouant sur la durée de trame (soit de la croche à la ronde par exemple); on pourrait obtenir un rapport plus important en modifiant le nombre de trames d'une syllabe, voire en jouant sur le nombre et la durée des trames. Pour des raisons de simplicité, nous nous sommes limités à la variation de la durée de trame dans les exemples proposés.

La "chanson" ainsi créée occupe le même nombre d'octets que la même expression non chantée; seuls les bits définissant la variation de hauteur et la durée de trame sont modifiés. Un exemple de programme permettant la création d'une expression chantée est donné pour deux des machines adaptées, et sa transposition aux autres à partir du programme de synthèse par phonèmes sera aisée.