ANALISI E SINTESI DEL SEGNALE VOCALE

di Leo Sorge

Che i computer parlino è una cosa talmente affascinante da lasciare ancora stupiti anche gli esperti del settore; ancora più incredibile è pensare che si può arrivare a ciò da tanti punti di partenza completamente scorrelati (elettronico, sistemistico, linguistico) da far capire che tutto sommato siamo ancora molto lontani dall'obiettivo pieno, anche se riusciamo a minimizzare gli errori quel tanto che basta a rendere comprensibile l'uscita di queste circuiterie. Tra l'altro, visto che il metodo usato dall'uomo per parlare è di tipo meccanico, viene da pensare che tutte queste soluzioni elettroniche siano soltanto una conseguenza dell'andamento della ricerca attuale, volta al microprocessore con annessi e connessi. Infatti così è, dato che nei secoli passati (le prime notizie attendibili sono della fine del '700) già qualcuno aveva scoperto i caratteri fondamentali del parlato, e li aveva riprodotti con i mezzi dell'epoca.

Fatto sta che l'aria modulata che esce dalla nostra bocca porta con sé un'informazione complessiva che potremmo memorizzare completamente con circa 300.000 bit al secondo: una cosa chiaramente impossibile da sfruttare utilmente per l'uomo della strada (che oggigiorno ha il suo home computer), tanto più che le informazioni effettive sono in quantità incredibilmente minore, nell'ordine dei 100 bit al secondo. Di conseguenza si deduce che i vari aspetti del segnale vocale sono grandemente interallacciati tra di loro, e che gran parte — la quasi totalità — non è fondamentale per la comprensione, come è stato dimostrato da quella che in ordine di tempo è l'ultimo ritrovato del settore, ovvero la sintesi per allofoni.

Da qui alla realizzazione di circuiti poco costosi, diciamo nell'ordine del prezzo di un accessorio per home computer, il passo è breve: e se il PCM, che è tanto bello, non fa per noi, consoliamoci almeno con i chip presenti sul mercato, che fanno un po' di casino ma tutto sommato si capiscono, e sono alla portata delle nostre tasche.

L'approccio elettronico

La prima conseguenza dell'avvento delle tecniche di elaborazione numerica fu la necessità di rendere tutte le grandezze elettriche, per loro natura continue, digeribili da sistemi basati su operazioni aritmetiche. La prima tecnica usata è il cosiddetto campionamento: a regolari intervalli di tempo si rileva il valore della grandezza e lo si approssima per difetto, ottenendo un numero (che può così essere elaborato da sistemi a microprocessore). La frequenza con cui gli intervalli di tempo si ripetono dipende dalla massima frequenza che si vuole riprodurre con questo sistema: il teorema di Shannon (o del campionamento) stabilisce che per poter riprodurre una frequenza di N cicli al secondo (o Hz) bisogna campionarla con una frequenza almeno pari a $2 \times N$.

La qualità, però, dipende anche da un secondo parametro, che è l'approssimazione che viene fatta su ogni campione: quello è tutto errore, che — trattandosi di segnali da riconvertire in grandezze acustiche — diventerà tutto rumore, quindi un ulteriore fattore di incomprensibilità. L'errore è dunque, per ogni campione, la differenza tra il valore reale e la sua approssimazione.

Un esempio di tecnica di conversione da segnale continuo (analogico) a digitale (numerico) e viceversa è mostrato in figura 1; questa tecnica è detta PCM, dalle iniziali di Pulse Code Modulation (ovvero modulazione secondo la codifica dell'impulso).

È evidente che questo metodo consente risultati molto buoni, dato che entrambi i parametri da cui deriva la qualità globale — frequenza di campionamento ed approssimazione — sono completamente sotto il controllo del progettista: la riproduzione di un segnale vocale con questo sistema porta ad un elevato grado di comprensione.

Il grande svantaggio del PCM è la quantità di memoria richiesta. Facciamo il conto: il segnale vocale emesso dall'uomo, pur avendo uno spettro più esteso, ha le sue componenti fondamentali nella gamma che giunge fino a circa 4000 Hz (cicli al secondo); la qualità della voce cui vengano tolte le componenti superiori a questa soglia è grosso modo paragonabile a quella del telefono, un po' chiusa e compressa. Per il teorema di Shannon avremo allora bisogno di una frequenza di campionamento pari a $2 \times 4000 = 8000$ cicli al secondo; se accetteremo solo 256 valori per l'ampiezza approssimata (nella fig. 1, per praticità ci siamo fermati a 7) per ogni campione avremo bisogno di un numero ad 8 bit. Tirando le somme, per ogni secondo di sintesi vocale avremo bisogno di 8000 campioni ad 8 bit, per un totale di 64000 bit/s: in soli 8 secondi riempiremmo ben 64K byte di RAM!

È evidente che ciò non è accettabile. Per mantenere la qualità intrinseca dei sistemi a campionamento senza dover ricorrere a intollerabili quantità di memoria, alcune case giapponesi ed americane hanno approntato migliorie al sistema, sostanzialmente basate su una tecnica di riduzione del numero di dati utili, detta Modulazione Logaritmica. Infatti qualunque numero può essere messo in forma più conveniente di quella lineare, usando un'approssimazione accettabile e molto conveniente per il risparmio di memoria. Per semplificare le cose, eccovi un esempio, che faremo con valori decimali in quanto più familiari di quelli binari usati dai computer:
il numero 1234567890 occupa 10 cifre;
mettendolo in forma logaritmica, con 4 cifre di mantissa, diventa
$0{,}1234 \times 10^{10}$,
che corrisponde alla seguente rappresentazione:
1234 10,
che occupa 6 cifre invece di 10, con un errore pari a 567890, in percentuale minore dell'1 per diecimila, e comunque sicuramente minore dell'errore di misura e campionamento. Si conclude che un errore ragionevole è il prezzo da pagare per ottenere un risparmio di spazio del 40%, cosa questa accettabilissima.

Nonostante esistano queste ed altre forme di compattazione dei dati (la correzione ad angolo di fase o Phase-Angle Adjustment, l'azzeramento a metà periodo o Half-period Zeroing), la quantità di memoria richiesta è sempre piuttosto elevata, e riduce le applicazioni comuni al campo dei secondi, al più di pochi minuti. Inoltre la natura del campionamento porta alla

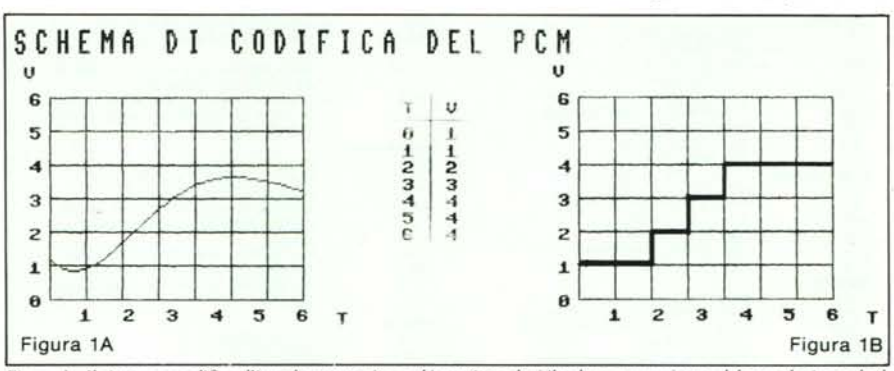

Figura 1 - Il sistema a codifica d'impulso memorizza ad istanti regolari il valore approssimato del segnale, in modo da ottenere un elenco di numeri che corrisponde abbastanza precisamente al segnale originale; è evidente che maggiore è la frequenza con cui si effettua (e memorizza) la misura, migliore è la precisione del sistema.
La ricostruzione avviene tenendo fisso il livello nell'intervallo tra due istanti di campionamento successivi; in questo modo si riproduce la forma d'onda originale, commettendo però un certo errore (la differenza tra l'onda originale e la sua approssimazione); tutto l'errore diventa disturbo udibile. L'opera viene completata da un filtro che smussa gli spigoli della forma d'onda, rendendola più digeribile al nostro orecchio.

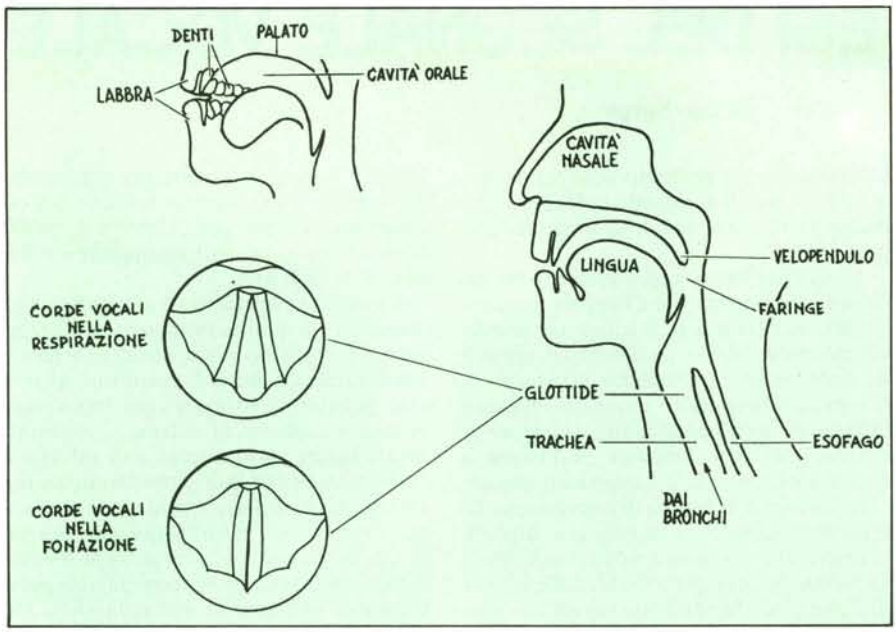

Figura 2: L'apparato fonatorio dell'uomo. L'aria espulsa dai polmoni viene ulteriormente compressa dalle corde vocali, due fasce muscolo-membranose che si chiudono nella prima fase dell'espirazione, per venir aperte dalla forza dell'aria stessa, che quindi esce più veloce. A questo punto una successione di condotti (bronchi, trachea, glottide, faringe) conduce il suono fino alla cavità orale, ove subisce l'ultima e più forte modulazione ad opera del palato, della lingua, delle labbra e dei denti. Anche le riflessioni della cavità nasale contribuiscono alla formazione del timbro della voce emessa.

cosiddetta sintesi per parole: il vocabolario è allora forzatamente limitato. E evidente che un approccio di questo tipo, prettamente elettronico e quindi totalmente privo di raffronti con la reale costruzione della voce, offre risultati praticamente insoddisfacenti, seppur qualitativamente elevati.

Gli approcci algoritmici

Il tipico sistema per non usare memoria è sfruttare i progressi dei vari settori della matematica attuale, cercando di individuare un procedimento o algoritmo che (a partire da eventuali dati iniziali) ricostruisca il segnale vocale tramite calcoli; in questo modo la memoria servirà solo per ospitare il procedimento matematico (che si ripeterà, identico, per ogni successivo valore calcolato) con un risparmio elevatissimo. Quella che rimane comunque alta è la quantità di bit al secondo trasmessi a quella parte del circuito che riconverte i segnali elettrici in onde sonore.

Silence	
PA1 (10 ms)	— before BB, DD, GG, and JH
PA2 (30 ms)	— before BB, DD, GG, and JH
PA3 (50 ms)	— before PP, TT, KK, and CH, and between words
PA4 (100 ms)	— between clauses and sentences
PA5 (200 ms)	— between clauses and sentences
Short Vowels	
*/IH/	— sitting, stranded
*/EH/	— extent, gentlemen
*/AE/	— extract, acting
*/UH/	— cookie, full
*/AO/	— talking, song
*/AX/	— lapel, instruct
*/AA/	— pottery, cotton
Long Vowels	
/IY/	— treat, people, penny
/EY/	— great, statement, tray
/AY/	— kite, sky, mighty
/OY/	— noise, toy, voice
/UW1/	— after clusters with YY: computer
/UW2/	— in monosyllabic words: two, food
/OW/	— zone, close, snow
/AW/	— sound, mouse, down
/EL/	— little, angle, gentlemen
R-Colored Vowels	
/ER1/	— letter, furniture, interrupt
/ER2/	— monosyllables: bird, fern, burn
/OR/	— fortune, adorn, store
/AR/	— farm, alarm, garment
/YR/	— hear, earring, irresponsible
/XR/	— hair, declare, stare
Resonants	
/WW/	— we, warrant, linquist
/RR1/	— initial position: read, write, x-ray
/RR2/	— initial clusters: brown, crane, grease
/LL/	— like, hello, steel
/YY1/	— clusters: cute, beauty, computer
/YY2/	— initial position: yes, yarn, yo-yo
Voiced Fricatives	
/VV/	— vest, prove, even
/CH1/	— word-initial position: this, then, they
/CH2/	— word-final and between vowels: bathe, bathing
/ZZ/	— zoo, phase
/ZH/	— beige, pleasure
Voiceless Fricatives	
*/FF/	—
*/TH/	— These may be doubled for initial position
*/SS/	— and used singly in final position
/SH/	— shirt, leash, nation
/HH1/	— before front vowels: YR, IY, IH, EY, EH, XR, AE
/HH2/	— before back vowels: UW, UH, OW, OY, AO, OR, AR
/WH/	— white, whim, twenty
Voiced Stops	
/BB1/	— final position: rib; between vowels: fibber; in clusters: bleed, brown
/BB2/	— initial position before a vowel: beast
/DD1/	— final position: played, end
/DD2/	— initial position: down; clusters: drain
/GG1/	— before high front vowels: YR, IY, IH, EY, EH, XR
/GG2/	— before high back vowels: UW, UH, OW, OY, AX; and clusters: green, glue
/GG3/	— before low vowels: AE, AW, AY, AR, AA, AO, OR, ER; and medial clusters: anger; and final position: peg
Voiceless Stops	
/PP/	— pleasure, ample, trip
/TT1/	— final clusters before SS: tests, its
/TT2/	— all other positions: test, street
/KK1/	— before front vowels: YR, IY, IH, EY, EH, XR, AY, AE, ER, AX; initial clusters: cute, clown, scream
/KK2/	— final position: speak; final clusters: task
/KK3/	— before back vowels: UW, UH, OW, OY, OR, AR, AO; initial clusters: crane, quick, clown, scream
Affricates	
/CH/	— church, feature
/JH/	— judge, injure
Nasal	
/MM/	— milk, alarm, ample
/NM1/	— before front and central vowels: YR, IY, IH, EY, EH, XR, AE, ER, AX, AW, AY, UW; final clusters: earn
/NN2/	— before back vowels: UH, OW, OY, OR, AR, AA
/NG/	— string, anger

*These allophones can be doubled.

Figura 3

Un primo modo algoritmico di ricostruire la voce umana è quello prettamente sistemistico, o ingegneristico: studiare la generazione originale (quella che avviene nell'uomo), scomporla in elementi singolarmente ottenibili tramite circuiti elettronici e infine ricomporla con i suddetti circuiti. Diamo un'occhiata all'apparato fonatorio del corpo umano: una certa circolazione d'aria viene indotta dai polmoni attraverso le corde vocali, due fasce muscolari che — dapprima chiuse — vengono poi aperte dalla forza dell'aria. Questa va a rimbalzare in una serie di condotti tra loro connessi (bronchi, trachea, esofago, glottide, faringe) fin quando giunge nella cavità orale, dove viene ulteriormente manipolata dalla lingua, dal palato, dai denti e dalle labbra (vedi fig. 2). Il percorso complessivo può esser realizzato tramite una successione di filtri in cascata (o con un unico filtro dalle caratteristiche equivalenti) che agiscono sulle grandezze elettriche raffiguranti il segnale vocale nello stesso modo in cui le varie cavità agiscono sull'aria forzata: il parametro principale di valutazione è la quantità (e il modo) d'aria che viene rifratta o assorbita dalle varie parti.

In questa maniera si ottiene un filtro a più elementi (descritto analiticamente da una funzione a 10-12 poli) che corredato di una circuiteria complementare genera un segnale qualitativamente accettabile con un limitato uso di memoria. Anche in questo caso, però, la sintesi è per parole, quindi il dizionario è limitato.

Un secondo approccio matematico è dato da un altro settore della matematica, l'analisi numerica, ed in particolare da un procedimento detto estrapolazione: si tratta di stabilire dalle forme d'onda dei segnali vocali, delle regole comportamentali tramite le quali sia possibile ricostruire la forma completa di un'onda come quella della voce, che assolutamente non è casuale bensì definita da precise regole. In questo modo da un numero limitato di valori iniziali si cerca di ricostruire l'intero inviluppo del segnale vocale.

L'approccio ad allofoni

Il problema principale rimane il numero di bit da trasmettere ogni secondo all'amplificatore: siamo sempre — per bene che vada — nell'ordine delle migliaia di bit/s. A qualcuno, però, venne in mente che ogni parola è composta da un numero limitato di simboli, presi tra quelli dell'alfabeto, che in inglese conta 26 simboli e in italiano 21: anche contando che esiste una pronuncia, per cui la nostra lingua ha 5 simboli vocalici e 7 suoni perché la 'e' o la 'o' possono essere aperte o chiuse, e che hanno la doppia pronuncia anche la "s" e la "z", e molte altre situazioni (la h è muta, ma cambia la pronuncia di c e g, che già da sole danno problemi davanti alle vocali, come mostrato in gallo e giallo; il gruppo /sc/, che corrisponde ad un unico suono, e così via), ed è evidente che se la pronuncia complica qualcosa nella nostra lingua, di gran lunga

maggiori devono essere i problemi che arreca all'inglese...

Per lingue non troppo complicate, quindi tipicamente quelle europee del ceppo latino (principalmente italiano, francese, spagnolo) e sassone (tedesco ed inglese), tutte le parole possono essere convertite nel loro equivalente fonetico — che prevede la distinzione tra le doppie c, g, e, o, s, z... — andando a pescare in un alfabeto di simboli detti grafemi, per i quali esiste una regola internazionale. Intanto osserviamo che l'informazione data dalla trascrizione

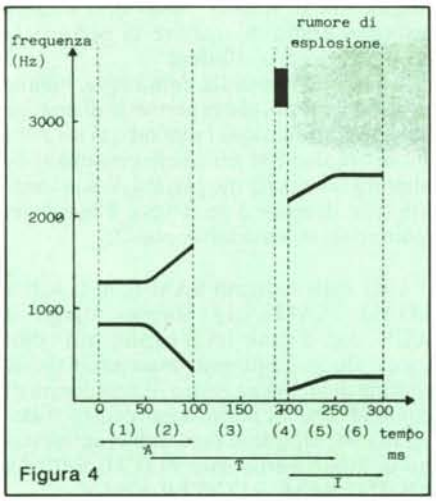

Figura 4

in grafemi non caratterizza completamente la parola, dato che nulla ci dice, ad esempio, sugli accenti in generale (un accento allunga la durata della vocale su cui cade) e su quelli di parole scritte nello stesso modo ma da leggersi con diversi accenti (si chiamano omografi), tipo càpitano (3a persona plurale del presente indicativo del verbo capitare), capitàno (1a persona del pres. ind. del verbo capitanare, ma anche sostantivo) e capitanò (3a persona singolare del passato remoto del verbo capitanare).

A questo punto sono stati proposti i fonemi: si tratta delle minime unità di suono del linguaggio, e comprendono gli accenti e la distinzione degli omografi. Ma anche i fonemi non bastano, perché la stessa lettera ha un suono differente se si trova all'inizio di una parola, in mezzo oppure alla sua fine: basti pensare alle parole molto, amico e tram, nelle quali lo stesso fonema, la /m/, ha tre diverse pronunce. Possiamo dunque definire l'allofono come l'unità minima di suono di un linguaggio, mentre il fonema è il simbolo di un gruppo di allofoni con qualcosa in comune; la figura 3 riporta fonemi ed allofoni della lingua inglese, cui sono orientati tutti i circuiti integrati in commercio.

Facciamo ancora un po' di conti sul numero di bit al secondo che dobbiamo trasmettere. Avendo a disposizione circa 60 allofoni (64 comprese diverse pause), per poterli distinguere ci serve un numero a 6 bit (111111 = 63); poiché il linguaggio parlato contiene da 10 a 12 elementi al secondo, siamo scesi a $6 \times 12 = 72$ bit al secondo, una quantità incredibilmente bassa se pensiamo ai 64000 del PCM prima maniera! Inoltre la sintesi per allofoni, basandosi sui mattoncini che compongono le parole, ha un vocabolario illimitato.

Questi due pregi fondamentali, che risolvono il problema di far parlare i computer per tutto ciò che è di basso costo come gli home computer, vanno a contrastare un grosso difetto: la qualità globale è piuttosto bassa, con un suono metallico e con elevato rumore. Perché ciò avviene? Studi sull'argomento hanno mostrato che la parola non si forma per allofoni. Il procedimento fisiologico porta invece alla costruzione per entità concatenate: in altre parole, il suono di una lettera non comincia mai quando finisce quello della precedente, bensì in uno stesso segmento troviamo elementi di più d'una componente (tipicamente 2). Ad indicare la veridicità di queste affermazioni si citano i risultati di una prova svolta: registrando la sillaba /ba/ non c'è modo di dividere il nastro in due parti tali che la prima suoni come /b/ e la seconda suoni come /a/, ma si sente prima un certo rumore e poi insieme il suono di /ba/. Questi esperimenti hanno confermato quindi due passi successivi: che la lingua non si propaga per allofono, ma al più per difoni (gruppi di due allofoni), e che le consonanti sono generate da fonti di rumore, mentre le vocali da suoni sinusoidali.

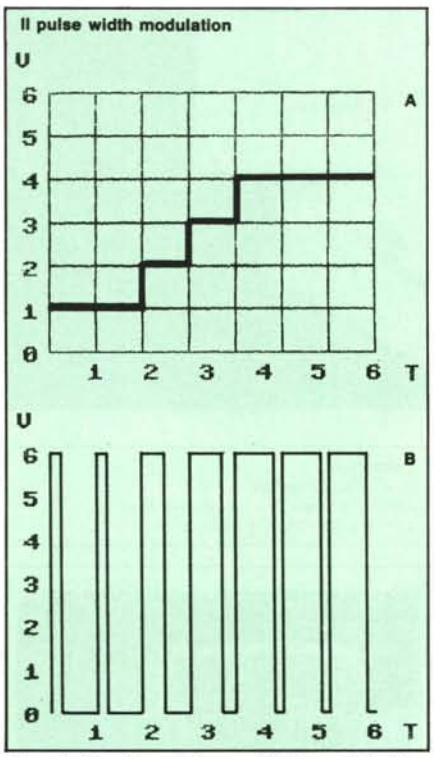

Figura 5 - Considerando la quantità di segnale (indipendentemente dalla sua forma) come unico parametro d'interesse è evidente che questa corrisponde all'area delle zone che si trovano sotto la linea a tratto continuo, che è già nella forma PCM. Per motivi circuitali è più semplice produrre in uscita forme d'onda che, invece di avere la base fissa e l'altezza variabile, abbiano sempre la medesima altezza ma basi variabili: variando la larghezza (width) dell'impulso si ottengono aree uguali. Un opportuno filtro sull'uscita ridarà alla forma PWM quella di un segnale continuo.

Se quindi il suono del parlato non si propaga a tratti consecutivi, è ovvio che una sua riduzione a questa forma eliminerà gran parte dell'informazione, che — come visto nel caso dell'errore di campionamento del PCM — si risolve in un certo rumore udibile, oltre che in una generale perdita di qualità, che comunque lascia il risultato perfettamente comprensibile nelle normali condizioni d'uso. La figura 4 mostra lo spettro della sillaba /ati/.

E in pratica?

Ciò che si trova nei negozi di elettronica è del tipo a fonemi, e si tratta di diversi chip dalle caratteristiche estremamente simili, come il General Instruments SP 0256 o il Digitalker della National. Sono inseriti nel contenitore standard da 40 piedini che vanno collegati da un lato al computer (con opportuno software di controllo) e dall'altro ad un amplificatore audio; il segnale d'uscita, per il quale si usa la modulazione della larghezza d'impulso (Pulse Width Modulation, vedi fig. 5), va opportunamente integrato prima di essere mandato all'amplificatore.

Più d'una rivista italiana di elettronica si è occupata di questo argomento, fornendo schemi di collegamento e di realizzazione: tra queste citiamo Elettronica 2000 Mister Kit, che nelle pgg. 46-50 del n° 60 - aprile '84 guida il lettore non alle prime armi alla realizzazione di una scheda di sintesi vocale per il VIC 20, basata sulle specifiche consigliate dalla General Instruments per il suo SP 0256. Il progetto è facilmente riconducibile al Commodore 64. La letteratura inglese offre un eccellente esempio sul numero 6, vol. 2, del periodico 'Electronics-The Maplin Magazine', che può esser richiesto alla stessa Maplin, al P.O. Box 3, Rayleigh, Essex 886 8LR, G.B., al prezzo di 1,30 sterline (circa) comprese spese postali: in questo numero troverete un articolo sulla sintesi per allofoni pubblicato con il permesso della stessa General Instruments più due progetti, uno per il VIC (collegabile al 64 con poche modifiche) e uno per lo ZX 81 (interfacciabile facilmente anche allo Spectrum).

L'interesse per l'argomento merita senz'altro un approfondimento: promettiamo di tornare prima possibile sulla sintesi vocale. Nel frattempo chi volesse maggiori ragguagli in generale può consultare l'articolo "Analisi, riconoscimento e sintesi del segnale vocale", di M. di Benedetto e G. Sommi in Note di Informatica dell'ottobre '83; per l'uso degli allofoni, invece, consigliamo "Allophone Speech Synthesis Technique", di Janet May in Electronics, anno II n° 6 (marzo-maggio '83).

Chi, possedendo un Commodore 64, volesse dei risvolti pratici, può leggere l'articolo pubblicato in questo stesso numero a pag. 102; un articolo sugli add-on parlanti per lo Spectrum è invece a pag. 68 del numero 29. MC