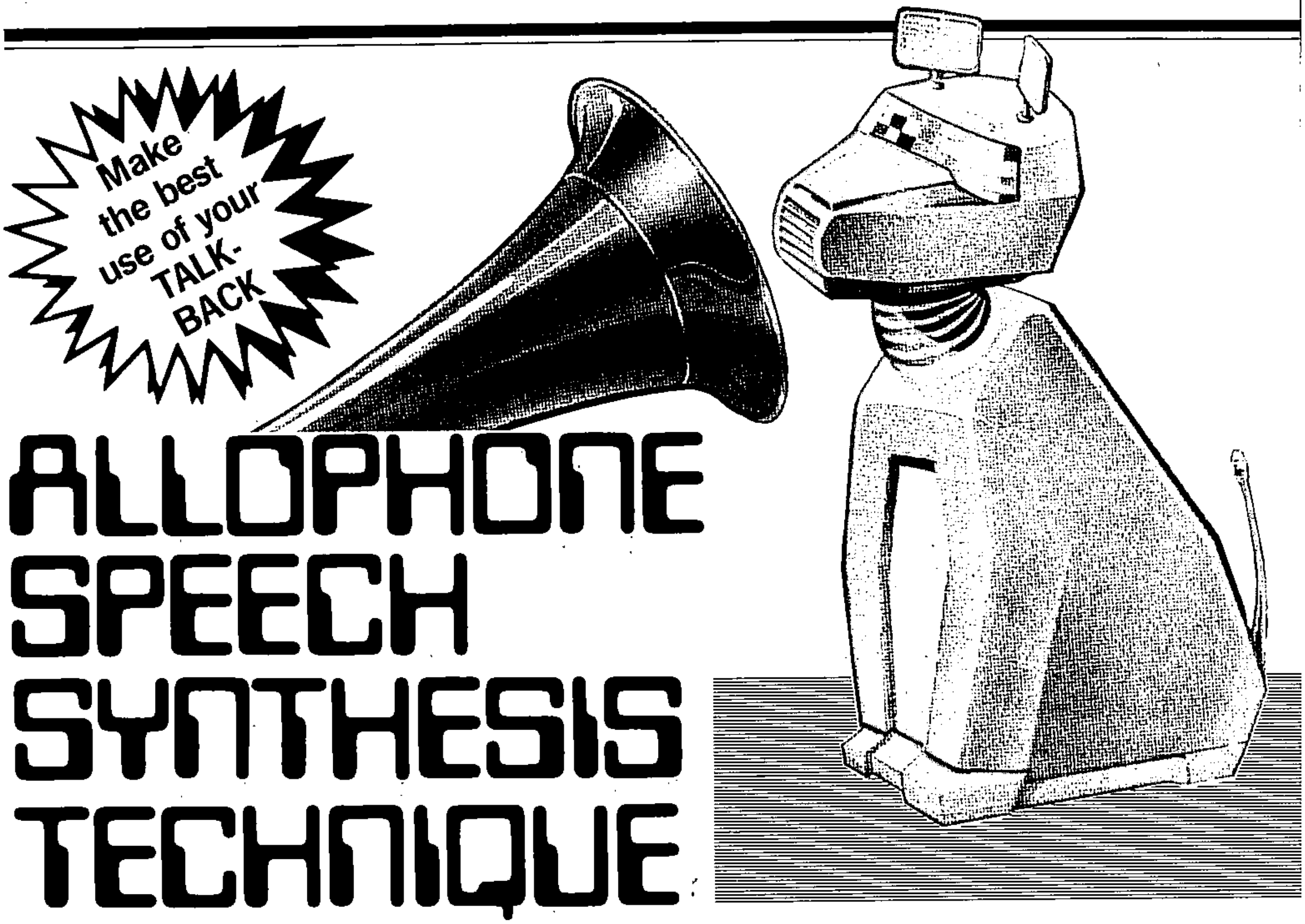

ALLOPHONE SPEECH SYNTHESIS TECHNIQUE

by Janet May

This article is reproduced by kind permission of General Instrument Microelectronics. Our many thanks to them.

Introduction

The General Instrument allophone speech synthesis technique is easy to use, has a remarkably low bit rate, and allows the user to synthesise any English word by concatenating individual speech sounds. Each allophone requires a six bit address. Assuming that speech contains ten to twelve allophones per second, allophone synthesis would require addressing less than 100 bits per second. Previous techniques have involved synthesising and storing entire words as units. The major disadvantage of this method is that, unless you want to use a very large memory, you are limited to a small vocabulary. For example, pulse code modulation (PCM), which is no more than digital recording, storage, and playback of speech waveforms, requires about 70 thousand data bits per second of speech. Another method, linear predictive coding (LPC), which predicts a speech sample from a weighted combination of previous samples, requires only one to two thousand bits per second to speech. Using this method, approximately 15-20 words can be stored in 16K bits of memory. While these methods require a large memory for a limited vocabulary, their big advantage is relatively high quality speech.

Allophone synthesis, on the other hand, has the major advantage of providing an unlimited vocabulary, since the stored units are not words, but individual speech sounds (allophones). The user merely has to become familiar with the speech sounds of English *(which are different from letters)* and the allophone symbols used to represent them. Another use for allophone synthesis is in a text-to-speech system in which the user inputs a string of text no different from what you are presently reading. The advantage of such a system is that the user does not have to learn the allophone symbols. Two sets of rules would be required: one which converts text to allophone symbols, and a second which converts those symbols to sounds. It is the second set of rules which we have already created and are discussing here.

	One-sound-to-many-letter representation	Many-sound-to-one-letter representation
Vowels	meat	vein
	feet	foreign
	Pete	deism
	people	deicer
	penny	geisha
Consonants	ship	although
	tension	ghastly
	precious	cough
	nation	

Table 1 - Spelling Irregularities

One disadvantage of allophone synthesis, however, is that, although completely understandable, the speech quality is not as good as it is for PCM or LPC. The problem arises when concatenating the allophones to form words. This will be discussed further in the sections to follow.

Language

In order to successfully use a set of allophone sounds to synthesise words there are a few preliminary points which should be made about speech and language. First, there is no one-to-one correspondence between written letters and the sounds of a language; secondly, speech sounds are not discrete units as beads on a string are; and lastly, speech sounds are acoustically different depending on what position in a word they occur, and what sounds precede or follow them.

The first of these is a problem which a child encounters when learning how to read. Each sound in a language may be represented by more than one letter, and conversely, each letter may represent more than one sound. (See the examples in Table 1). Because of these spelling irregularities we must be very careful to remember to think in terms of *sounds* not letters, when dealing with speech.

The second point to be made concerns segmentation of the speech signal. An adult who has learned how to read usually thinks of the acoustic stream of speech as a string of discrete sounds which he calls by their letter names. But, in fact, speech is a continuously varying signal which cannot be easily broken into distinct sound-size units. For example, if one attempts to extract the b sound from the word bat by taking successively larger chunks of the acoustic signal from the beginning of the word, one at first hears a non-speech noise, and then at some point hears ba. In other words, there is no point at which the b sound can be heard in isolation; one hears either a non-speech noise or the syllable ba.

Finally, the most important point to make for users of an allophone set, is that the acoustic signal of a speech sound may differ depending on whether it occurs in word-initial or word-final position; or in the environment of a vowel which is articulated in the front or back of the oral cavity, a long or short vowel, or a voiced or voiceless consonant. For example, the initial p in pop will be acoustically different from the p in spy,

		Labial	Labio-Dental	Inter-Dental	Alveo-Lar	Palatal	Velar	Glottal
Stop:	Voiceless	PP			TT		KK	
	Voiced	BB			DD		GG	
Fricatives:	Voiceless	WH	FF	TH	SS	SH		HH
	Voiced		VV	DH	ZZ	ZH*		
Affricates:	Voiceless					CH		
	Voiced					JH		
Nasals:	(Voiced)	MM			NN		NG*	
Resonants:	(Voiced)	WW			RR,LL	YY		

Labial: Upper and Lower Lips Touch or Approximate
Labio-Dental: Upper Teeth and Lower Lip Touch
Inter-Dental: Tongue Between Teeth
Alveolar: Tip of Tongue Touches or Approximates Alveolar Ridge (just behind upper teeth)
Palatal: Body of Tongue Approximates Palate (roof of mouth)
Velar: Body of Tongue Touches Velum (posterior portion of roof of mouth)
Glottal: Glottis (opening between vocal cords)

*These do not occur in word-initial position in English.
**Examples of these phonemes in word context can be found in Table 5.

Table 2 - Consonant Phonemes of English**

	Front	Central	Back
High	YR		
	IY		UW#
	IH*		UH*#
Mid	EY	ER	OW#
	EH*	AX*	OY#
	XR		
Low	AE*	AW#	AO*#
		AY	OR#
*Short Vowels		AR	
#Rounded Vowels		AA*	

Table 3 - Vowel Phonemes of English

and may be different from the final p in pop. Furthermore, the ear will perceive the same acoustic signal differently, depending on what sounds precede or follow it. The word cot can be made to sound like cod by lengthening the duration of the O, and conversely, the word cod can be made to sound like cot by shortening the duration of the O.

Phonemes of English

It will be useful to know what the speech sounds of English are. The sounds of a language are called phonemes, and each language has a set of which is slightly different from those of other languages.

	(Silence)			(Voiced Fricat.)	
PA1	PAUSE	10MS	/VV/	vEST	190MS
PA2	PAUSE	30MS	/DH1/	thEY	290MS
PA3	PAUSE	50MS	/DH2/	thEY	120MS
PA4	PAUSE	100MS	/ZZ/	zOO	210MS
PA5	PAUSE	200MS	/ZH/	AzURE	190MS
	(Short Vowels)			(Voiceless Fricat.)	
x/IH/	SiT	70MS	x/FF/	fOOD	150MS
x/EH/	eND	70MS	x/TH/	thIN	180MS
x/AE/	HaT	120MS	x/SS/	VEsT	90MS
x/UH/	BooK	100MS	/SH/	shIP	160MS
x/AO/	auGHT	100MS	/HH1/	hE	130MS
x/AX/	SuCCEED	70MS	/HH2/	hOE	180MS
x/AA/	HoT	100MS	/WH/	whIG	200MS
	(Long Vowels)			(Voiced Stop Cons.)	
/IY/	See	250MS	/BB1/	bUSINESS	50MS (SOFT)
/EY/	BeiGE	280MS	/BB2/	bUSINESS	50MS
/AY/	SKy	260MS	/DD1/	COULd	70MS
/OY/	Boy	420MS	/DD2/	dO	160MS
/UW1/	To	100MS	/GG1/	gUEST	80MS
/UW2/	To	260MS	/GG2/	gOT	50MS
/OW/	Beau	240MS	/GG3/	WIg	160MS
/AW/	ouT	370MS		(Voiceless Stop Cons.)	
/EL/	SADDle	190MS	/PP/	pOW	210MS
	(R — Coloured Vowels)		/TT1/	PARt	100MS
/ER1/	Fir	160MS	/TT2/	tO	140MS
/ER2/	Fir	300MS	/KK1/	cAN'T	160MS
/OR/	STore	330MS	/KK2/	SkY	190MS
/AR/	ALarM	290MS	/KK3/	cOMB	120MS
/YR/	CLear	350MS		(Affricate)	
/XR/	REPair	360MS	/CH/	chURCH	190MS
	(Resonants)		/JH/	DOdgE	140MS
/WW/	wOOL	180MS		(Nasal)	
/RR1/	rURAL	170MS	/MM/	mILK	180MS
/RR2/	BrAIN	120MS	/NN1/	THIn	140MS
/LL/	lAKE	110MS	/NN2/	nO	190MS
/YY1/	yES	130MS	/NG/	AnCHOR	220MS
/YY2/	yES	180MS			

x — These allophones can be doubled

Table 4. Allophones

Table 2 contains a chart of all the consonant phonemes of English, and Table 3 all the vowel phonemes of English.

Consonants are produced by creating a constriction or complete occlusion in the vocal tract which produces an aperiodic sound source. If the vocal cords are vibrating at the same time, as in the case of the voiced fricatives VV, DH, ZZ, and ZH (see Table 4) there are two sound sources: one which is aperiodic and one which is periodic.

Vowels are produced with a relatively open vocal tract and a periodic sound source (unless they are whispered) provided by the vibrating vocal cords. Vowels are classified according to whether the front or back of the tongue is high or low (see Table 3), whether they are long or short, and whether the lips are rounded or unrounded. In English all rounded vowels are produced in or near the back of the mouth (UW, UH, OW, AO, OR, AW).

It will be useful to remember that sounds which have features in common behave in similar ways. For example, the voiceless stop consonants PP, TT, and KK (see Table 2) require 50-80 msec of silence before them and the voiced stop consonants BB, DD and GG require 10-30 msec of silence before them. When you find a particular technique that works well with one sound, try using that same technique with similar sounds. For example, if you decide that KK1 sounds good before a front vowel (IY), use it before other front vowels (YR, IY, IH, EY, EH, XR, AE).

Allophones

So far we have been talking about phonemes, but in fact, a phoneme is an abstraction. It is the name given to a group of similar sounds in a language. Recall the statement that the phoneme PP will be acoustically different depending on whether it occurs in word-initial or word-final position, or after SS. Each of these different PPs are allophones of the phoneme PP. An allophone, therefore, is what occurs in the actual acoustic speech signal. A phoneme is the name of a group of related allophones. It is for this reason that our inventory of English speech sounds is called an allophone set.

How to use the allophone set

The allophone set (see Table 4) contains two or three versions of some phonemes. You may find that you need to use one allophone or a particular phoneme for word - or syllable - initial position and another for word - or syllable - final position. A detailed set of guidelines for using the allophones is given in Table 6. Note that these are suggestions, not rules.

DD2-AO-TT2-ER1	"daughter"
KK3-AX1-LL-AY-DD1	"collide"
SS-SS-IH-SS-TT2-ER1	"sister"
KK1-LL-AW-NN1	"clown"
SS-KK3-WW-XR	"square"
KK3-UH-KK1-IY	"cookie"
LL-EH-TT2-ER	"letter"
LL-IH-TT2-EL	"little"
AX1-NG-KK3-EL	"uncle"
KK1-AX1-MM-PP1-YY1-UW1-TT2-ER	"computer"
EH-KK1-SS-TT2-EH-EH-NN1-TT2	"extent"
TT2-UW2	"two"
AX1-LL-AR-MM	"alarm"
SS-KK3-CR	"score"
FF-ER2	"fir"

Table 5 - Examples of words made from Allophones

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/ —
*/SS/ —
These may be doubled for initial position 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.

Table 6. Guidelines for using the Allophones.

Decimal Address	Octal Address	Hex Address	Allophones	Sample Word	Duration
0	000	0	PA1	PAUSE	10MS
1	001	1	PA2	PAUSE	30MS
2	002	2	PA3	PAUSE	50MS
3	003	3	PA4	PAUSE	100MS
4	004	4	PA5	PAUSE	200MS
5	005	5	/OY/	Boy OY	420MS
6	006	6	/AY/	Sky Y	250MS
7	007	7	/EH/	End E	70MS
8	010	8	/KK3/	Comb C	120MS
9	011	9	/PP/	Pow P	210MS
10	012	A	/JH/	Dodge G	140MS
11	013	B	/NN1/	Thin N	140MS
12	014	C	/1H/	Sit I	70MS
13	015	D	/TT2/	To T	140MS
14	016	E	/RR1/	Rural R	170MS
15	017	F	/AX/	Succeed U	70MS
16	020	10	/MM/	Milk M	180MS
17	021	11	/TT1/	Part T	100MS
18	022	12	/DH1/	They TH	290MS
19	023	13	/IY/	See E	250MS
20	024	14	/EY/	Beige EI	280MS
21	025	15	/DD1/	Could ID	70MS
22	026	16	/UW1/	To O	100MS
23	027	17	/AO/	Aught AU	100MS
24	030	18	/AA/	Hot O	100MS
25	031	19	/YY2/	Yes YE	180MS
26	032	1A	/AE/	Hat A	120MS
27	033	1B	/HH1/	He H	130MS
28	034	1C	/BB/	Business BU	80MS
29	035	1D	/TH/	Thin TH	180MS
30	036	1E	/UH/	Book OO	100MS
31	037	1F	/UW2/	Food OO	260MS
32	040	20	/AW/	Out OU	370MS
33	041	21	/DD2/	Do D	160MS
34	042	22	/GG3/	Wig IG	140MS
35	043	23	/VV/	Vest V	190MS
36	044	24	/EG1/	Guest GU	80MS
37	045	25	/SH/	Ship S	160MS
38	046	26	/ZH/	Azure Z	190MS
39	047	27	/RR2/	Brain R	120MS
40	050	28	/FF/	Fdod F	150MS
41	051	29	/KK2/	Sky K	190MS
42	052	2A	/KK1/	Can't C	160MS
43	053	2B	/ZZ/	Zoo Z	210MS
44	054	2C	/NG/	Anchor N	220MS
45	055	2D	/LL/	Lake L	110MS
46	056	2E	/WW/	Wool W	180MS
47	057	2F	/XR/	Repair R	360MS
48	060	30	/WH/	Whig W	200MS
49	061	31	/YY1/	Yes Y	130MS
50	062	32	/CH/	Church C	190MS
51	063	33	/ER1/	Fir IR	160MS
52	064	34	/ER2/	Fir ERR	300MS
53	065	35	/CW/	Beau AU	240MS
54	066	36	/DH2/	They TH	240MS
55	067	37	/SS/	Vest S	90MS
56	070	38	/NN2/	No N	190MS
57	071	39	/HH2/	Hoe H	180MS
58	072	3A	/OR/	Store OR	330MS
59	073	3B	/AR/	Alarm A	290MS
60	074	3C	/YR/	Clear R	350MS
61	075	3D	/EG2/	Got G	40MS
62	076	3E	/EL/	Saddle L	190MS
63	077	3F	/BB2/	Business B	50MS

Allophone Address Table.

For example, DD2 sounds good in initial position and DD1 sounds good in final position, as in "daughter" and "collide". (See Table 5 for instructions on how to create all the sample words mentioned in this section). One of the differences between the initial and final versions of a consonant is that an initial version may be longer than the final version. Therefore, to create an initial SS, you can use two SSs instead of the usual single SS at the end of a word or syllable, as in "sister". Note that this can be done with TH, and FF, and the inherently short vowels (to be discussed below), but with no other consonants. You will want to experiment with some consonant clusters (strings of consonants such as str, cl) to discover which version works best in the cluster. For example KK1 sounds good before LL as in "clown", and KK2 sounds good before WW as in "square". One allophone of a particular phoneme may sound better before or after back vowels and another before or after front vowels. KK3 sounds good before UH and KK1 sounds good before 1y, as in "cookie". Some sounds (PP, BB, TT, DD, KK, GG, CH and JH) require a brief duration of silence before them. For most of these, the silence has already been added but you may decide you want to add more. Therefore, there are several pauses included in the allophone set varying from 10-200 msec. To create the final sounds in the words "letter" and "little" use the allophones ER and EL. Remember that you must always think about how a word sounds, not how it is spelled. For example, the NG allophone obviously belongs at the ends of the words "sing" and "long", but notice that the NG sound is represented by the letter N in "uncle". And remember that some sounds may not even be represented in words by any letters, as the YY in "computer".

As mentioned earlier there are some vowels which can be doubled to make longer versions for stressed syllables. These are the inherently short vowels IH, EH, AE, AX, AA and UH. For example, in the word "extent" use one EH in the first syllable, which is unstressed and two EHs in the second syllable which is stressed. Of the inherently long vowels there is one, UW, which has a long and short version. The short one, UW1, sounds good after YY in computer. The long version, UW2, sounds good in monosyllabic words like "two". Included in the vowel set is a group called R-coloured vowels. These are vowel + R combinations. For example, the AR in "alarm" and the OR in "score". Of the R-colored vowels there is one, ER, which has a long and short version. The short version is good for polysyllabic words with final ER sounds like "letter", and the long version is good for monosyllabic words like "fir". One final suggestion is that you may want to add a pause of 30-50 msec between word, when creating sentences, and a pause of 100-200 msec between clauses.