Computer-coding the IPA: a proposed extension of SAMPA

J.C.Wells, University College London

1. **Computer coding**. When an ASCII file (a DOS text file) is sent as an e-mail message, the only characters that are sure to be properly transmitted are those with ASCII/ANSI numbers between 32 and 126. These comprise upper-case A..Z, lower-case a..z, numerals 0..9, punctuation marks ! " ' ( ) , - . / : ; ? [ ] { }, other marks # $ % & * + < = > @ \ ^ _ ` | ~, and space. If we want to transmit phonetic symbols, we must therefore recode them using only these characters.

It is not even the case that all the 'other marks' mentioned will necessarily reappear correctly at the receiving end: on my own British screen, for example, an incoming character originally transmitted as a hash mark (#) appears as pound sterling (£). But at least there is a consistent one-for-one substitution, so that information is not lost. On the other hand an outgoing character falling outside the range 32..126 is very likely to be converted into something else: a pound sterling sign (£, ASCII 156, ANSI 0163) transmitted from a British keyboard may be received (even in the UK) as an exclamation mark (!), hash (#), or other substitute.

2. The **SAM Phonetic Alphabet** (SAMPA) conventions were drawn up in 1988-1991 (with subsequent minor revisions and extensions) by the SAM (Speech Assessment Methods) consortium, comprising speech scientists from nine countries of the European Community (Wells et al, 1992). The purpose of SAMPA was to form the basis of an international standard machine-readable phonetic alphabet for purposes of international collaboration in speech research. SAMPA as presently constituted covers all the symbols needed for the phonemic transcription of the principal European Union languages. One useful application is for sending phonetic symbols by e-mail.

As well as **codings** for symbols of the International Phonetic Alphabet (IPA), the SAMPA recommendations also cover choice of **transcription** for the European languages involved.

A convenient tabulation of the IPA symbols approved by the International Phonetic Association is the IPA Chart, the current edition of which is dated 1993 (IPA, 1993).

3. The IPA phonetic symbols that are also **lower-case alphabet** symbols naturally remain the same in SAMPA. These are the lower-case Latin letters a..z, ASCII/ANSI 97..122. SAMPA recodes all other phonetic symbols covered within the range 37..126. The present SAMPA recommendations are set out in the following table. To transmit the IPA character shown in the first column and described in the second, the user would have to type the standard character in the third column, which has the ASCII/ANSI code number in the fourth column.

IPA	***Comments***	***SAMPA***	***ASCII/ANSI***
Vowels			
ɑ	script a, open back unrounded vowel, card. 5, Eng. *start*	A	65
æ	ae ligature, near-open front unrounded vowel, Eng. *trap*	{	123
ɐ	turned a, open schwa, Ger. *besser*	6	54
ɒ	turned script a, open back rounded vowel, Eng. *lot*	Q	81
ɛ	epsilon, open-mid front unr.vowel, card. 3, Fr. *même*	E	69

ə	turned e, schwa, Eng. *banana*	@	64
ɜ	reversed epsilon, long mid central vowel, Eng. *nurse*		3
51			
ɪ	small cap I, lax close front unrounded vowel, Eng. *kit*	I	73
ɔ	turned c, open-mid back rounded vowel, Eng. *thought*	O	79
ø	slashed o, close-mid front rounded vowel, Fr *deux*	2	50
œ	oe ligature, open-mid front rounded vowel, Fr *neuf*	9	57
Œ	capital OE ligature, open fr. rounded V, Dan. *drømme*	&	38
ʊ	upsilon, lax close back rounded vowel, Eng. *foot*	U	85
ʉ	barred u, close central rounded vowel, Swed. *sju*	}	125
ʌ	turned v, open-mid back unrounded vowel, Eng. *strut*	V	86
ʏ	small cap Y, lax close front rounded vowel, Ger, *hübsch*	Y	89

Consonants

β	beta, voiced bilabial fricative, Sp. *cabo*	B	66
ç	c-cedilla, voiceless palatal fricative, Ger. *ich*	C	67
ð	edh, voiced dental fricative, Eng. *then*	D	68
ɣ	gamma, voiced velar fricative, Sp. *fuego*, Gk. *gama*	G	71
ʎ	turned y, palatal lateral, It. *famiglia*	L	76
ɲ	left-tail n, palatal nasal, Sp. *año*	J	74
ŋ	eng, velar nasal, Eng. *thing*	N	78
ʁ	inverted small cap R, voiced uvular (see below), Fr. *roi*	R	82
ʃ	esh, voiceless palatoalveolar fricative, Eng. *ship*	S	83
θ	theta, voiceless dental fricative, Eng. *thin*	T	84
ɥ	turned h, labial-palatal semivowel, Fr. *huit*	H	72
ʒ	ezh (yogh), voiced palatoalveolar fric., Eng. *measure*	Z	90
ʔ	glottal stop, German *Verein*, also Danish stød	?	63

Length, stress and tone marks

ː	length mark	:	58
ˈ	primary stress		"
34			
ˌ	secondary stress	%	37
˩	falling tone *(but see below)*	‘	96
˧	rising tone *(but see below)*	’	39

Diacritics (shown with another symbol as an example)

n̩	syllabic consonant, inferior stroke, Eng. *garden*	n=	60
ɔ̃	nasalization, superior tilde, Fr. *bon*	O~	126

4. This still leaves out many phonetic symbols we may wish to transmit. Accordingly, I now

propose one minor revision and then a two-stage **extension of the SAMPA conventions.** The revision and first stage of extension will provide for a number of relatively frequently used symbols, which it would be convenient to access directly from the PC keyboard. These include the remaining Cardinal Vowel symbols, together with other symbols needed for the phonemic transcription of Russian, Chinese, and Japanese, as well as of Swedish, Welsh, and American English, and symbols for important English allophones or variants. The second stage of the extension will provide SAMPA conventions to cover **all** remaining IPA symbols.

5. With SAMPA in its current form there is a problem in transcribing Russian and various other languages in which consonant palatalization plays an important role. This applies not only to several other languages of Eastern Europe, but also, for instance, to Irish. Russian has a phonemic opposition between pairs of consonants, one **palatalized** ('soft') and the other not, e.g. /brat^{j}/ брать 'to bring' vs. брат /brat/ 'brother'. Currently, the only way to write the palatalized consonants in SAMPA is as a two-place symbol, consonant plus [j]. This is unsatisfactory, particularly since the language also has sequences of non-palatalized consonant plus a separate /j/.

The IPA now recommends that palatalization be shown by a raised [j] after the consonant symbol in question, thus [^{j}]. Linguists working on Slavonic languages often prefer an acute accent accent, thus /brat´/. This convention is also used by scholars of Irish. I propose that SAMPA combine these two conventions by **redefining the apostrophe symbol**, ASCII 39, as a **palatalization** diacritic, the equivalent of IPA [^{j}].

In the current SAMPA ASCII 39 stands for Rising tone. This recommendation has not in practice been popular. Most transcribers and labellers have not used it, since they restrict themselves to segmental symbols plus stress. Gibbon, Hirst, and Barry have devised a prosodic extension of SAMPA, SAMPROSA, of which more below. Among the SAMPROSA proposals was one that R and / be adopted as alternatives to this use of ASCII 39. The colleagues mentioned have confirmed that they would have no objection to SAMPA redefining the apostrophe.

Thus we should have a new SAMPA diacritic:

t^{j}	raised j, palatalized	t’	39

(Depending on font and printer, ASCII 39 may appear on screen and printout not as a raised-comma apostrophe ’, but as an acute accent ´ or a vertical typewriter-style apostrophe '.)

With this revision it becomes possible to transcribe Russian and other Slavonic languages directly from the keyboard, as SAMPA /ak"t’abr’/, IPA /ak$^{\text{ˈ}}$t^{j}abr^{j}/ октябрь. It would also incidentally probably provide an acceptable notation for Polish *s, z*, IPA /ɕ, ʑ/, which could become SAMPA /s’, z’/.

Some other computer-readable alphabets use the apostrophe as the stress mark, with our double inverted commas as the secondary stress mark. But I don't see a strong case for changing our current recommendation of ["] *for primary and* [%] *for secondary stress.*

6. Redefinition of ASCII 39 suggests that we might extend parallel treatment to the grave accent, or reverse apostrophe, ‘ or `, ASCII 96. Hitherto SAMPA has defined it as meaning Falling tone, but for this the SAMPROSA group suggest F or \ as alternatives. I propose that it be given the role of **diacritic to show r-colouring**, on the current IPA chart termed Rhoticity. This is a feature of the General American vowels of *color* and *bird*, in the Kenyon and Knott transcription /ˈkʌlɚ/, /bɝd/. Under this proposal SAMPA would represent these as /"kVl@`/ and /b3`d/.

ɚ	r-coloured	@`	96

Actually, this diacritic could well be given a wider interpretation. The IPA symbols for retroflex consonants, [ʈ, ɖ, ɳ, ʂ, ʐ, ɭ, ɽ, ɻ], do not formally contain a diacritic, although they all include a rightward tail. It is a short step to propose that ASCII 96 be used in SAMPA to show retroflex consonants, by attaching it to the symbol for the corresponding alveolar. No ambiguity arises through this: the IPA rhoticity hook applies to symbols for vowels, the retroflexion tail to consonants. In any case, r-coloured ('rhotic') vowels are sometimes themselves termed retroflex.

ʈ	right-tail t, voiceless retroflex plosive, Hindi *t*	t`	96

— and likewise for ʈ, ɖ, ɳ, ʂ, ʐ, ɭ, ɽ (d`, n`, s`, z`, l`, r`). For ɻ, see below.

In Swedish the retroflex consonants can convincingly be phonemicized as /rC/, e.g. [kuʈ] /kurt/ *kort*. It has therefore been agreed that in SAMPA we write Swedish retroflex [ʈ] as "rt", [ɳ] as "rn", etc. But this solution is not satisfactory for languages in which a retroflex consonant is phonologically distinct from the sequence of /r/ and the corresponding alveolar/dental consonant, as is the case in several languages of India.

7. The **Polish** vowel spelt *y* is a close central vowel, IPA [ɨ]. Of the keyboard symbols not so far preempted for other purposes, the most suitable one to represent this vowel seems to be the numeral [1] (ASCII 49), which has some visual resemblance. Thus we would transcribe Polish *ryby* 'fish' as /r1b1/.

For the rounded mid central short *u* of Swedish, IPA [ɵ], SAMPA currently prescribes [u0]. This is an unsatisfactory ad-hoc retranscription rather than a proper coding. I suggest we change it another unoccupied keyboard character, the numeral [8], which has some visual similarity to [ɵ].

Two of the Cardinal Vowels have as yet no SAMPA version: IPA [ɯ, ɤ]. I suggest that [ɯ] be coded [M], on grounds of visual similarity, and [ɤ] [7], since it is secondary cardinal no. 7.

8. This leaves the following alphabetic keyboard symbols (ASCII 33..126) **unallocated** in SAMPA: 0 *(zero)*, 4, 5, F, K, P, W, X. Preliminary soundings among colleagues suggest that they should be distributed among the following IPA symbols, seen as most widely useful of those not yet provided for (excluding from consideration diacritics, which we turn to in due course):

[ɾ], alveolar tap (though in many languages it can be phonemicized as /r/)
[ɬ], dark l, an important allophone in English and Portuguese
[ɱ], labiodental nasal, an important allophone of /m/ in many languages

[ɬ], voiceless alveolar lateral fricative, Welsh *ll*, a lisped [s]
[ʋ], labiodental approximant, social variant of English /r/, one variant of Dutch /w/
[ʍ], voiceless labial-velar fricative, one version of English *wh*
[χ], voiceless uvular fricative, variant of /x/ in several languages

I suggest that the zero be reserved for use as a diacritic. We could assign the remaining symbols in the order shown: IPA [ɾ] becomes SAMPA [4], [ɬ] becomes [5], [ɱ] becomes [F] (in favour of which is the association of F with labiodental place), [ɬ̠] becomes [K], [ʋ] becomes [P], [ʍ] becomes [W], and [χ] becomes [X] (the last two being obviously appropriate).

9. Summarizing, these proposals are:

Vowels

ɨ	barred i, close central unr. vowel, Russ. ты, Welsh *tu*	1	49
ɵ	barred o, close-mid central rounded vowel, Swed. *buss*	8	56
ɯ	turned m, close back unrounded vowel, Korean *u*	M	77
ɤ	ram's horns, close-mid back unrounded vowel, card. 15	7	55

Consonants

ɬ	belted l, voiceless alveolar lateral fricative, Welsh *llaw*	K	75
ɫ	tilde-thru-l, dark l, Eng. *milk*	5	53
ɱ	tailed m, labiodental nasal, Eng. *emphasis*	F	70
ɾ	fish-hook r, voiced alveolar tap	4	52
ʋ	cursive v, labiodental approximant	P	80
ʍ	turned w, voiceless labial-velar fricative, Scottish *when*	W	87
χ	chi, voiceless uvular fricative, Welsh *bach*	X	88

Diacritic and pseudodiacritic (shown with another symbol as an example)

ɚ	diacritic for r-colouring, American Eng. *better*	@`	96
ʈ	diacritic for retroflex, Hindi *roti*	t`	96

Of the ASCII/ANSI range 33..126 we have now accounted for everything except the punctuation (etc.) marks . ! ( ) , _ . / ; < > [ ] ^ $ * + \ | #. It seems best not to recode these as phonetic symbols: we may well want to reserve them for their normal functions, including possible use as indicators of pauses and boundaries in transcription (for which both SAMPA/SAMPROSA and the IPA also make a number of recommendations).

The UK PC keyboard also usually includes the characters £ and ¬ (right-corner, extended ASCII 170, ANSI 0172); but these characters are not available on American keyboards and fall outside the range reliably usable for e-mail and similar purposes (see above). PC keyboards in other countries often include further characters not available on the US or UK keyboard, such as é (e-acute), ñ (n-tilde), ß (eszet), § (section-mark) — but these likewise are characters outside the range 33.126, and SAMPA does not make use of them.

10. The IPA has always allowed not only a **variety** of phonemic solutions, but also a variety of

transcriptional possibilities for any given language. We must always consider the possibility of solving phonetic symbol difficulties by turning to retranscription. In applying the IPA to Italian and Spanish many phoneticians follow the common-sense rule of writing the alveolar trill as [rr] and the tap as [r], rather than using [r] in its narrower meaning of trill, which necessitates the use of the relatively exotic symbol [ɾ] for the tap. This is SAMPA practice, too, for these languages, extended indeed in Spanish to writing the palatal fricative as [jj] as against single [j] for the palatal approximant; and thus SAMPA has managed so far without defining a coding for IPA [ɾ̠, j̠].

We may conclude that in principle when designing a machine-readable, e-mailable version of the IPA we need to be able to encode every symbol on the IPA chart. We may at present be focussing particularly on languages of Eastern Europe and the Far East, but in principle this should mean that we should render the extended SAMPA adequate to transcribe any language in the world.

11. In fact, with the extensions proposed up to this point, together with appropriate decisions on phonemicization, SAMPA will make it possible to handle **almost all the languages** on which speech researchers are currently working or likely to work. It will cover the segmental phonemics not only of the languages of the EU, but also of Russian, Hindi/Urdu, Chinese, Japanese, and many other languages besides.

It is, however, not yet complete. Our discussion so far has identified reasonable coding solutions for all the more widely-used symbols; the remainder must be dealt with in other ways. Major gaps include ħ and ʕ, which have phonemic status in Arabic; ɹ, a narrower symbol for the phonemic norm of English /r/; ɦ, an allophone of English /h/; and the many details of articulation that the IPA represents by the use of diacritics.

What then of the IPA symbols not yet provided for? One possibility would be to use the "IPA numbers" from the scheme established in 1989, recently reviewed by Esling and Gaylord, 1993. Or there are use other numerical codes we might use: the Unicode and the 40,000-strong Universal Character Set (UCS) defined by ISO 10646, or the Association for Font Information Interchange (AFII) codes (ibid). However, such numerical methods hardly offer a user-friendly or readable solution to the problem. We must cast around further. I have a number of proposals to put forward — the second stage in extending SAMPA.

12. **The backslash**. In deference to the requirements of those working on French, SAMPA has defined a special role for the forward slash, / (ASCII 47), namely as a marker of certain vowel archiphonemes or indeterminacies, e.g. *maison* /mE/zO~/. It is of course also widely used as a delimiter of phonemic transcriptions. However no segmental role is currently defined for the backslash (\, ASCII/ANSI 92), though there is a vague proposal that it could be used for "phonetic case shift". I propose that we now define this shift very precisely and bring the backslash into service as a kind of universal diacritic, meaning 'the preceding character is to be interpreted in a special way'. IPA alphabetic characters not yet assigned a SAMPA code can be given a two-place code, in which the second character is \.

Used in this way, the backslash will double the number of symbols we can accommodate within codes 33..127. The syntax must be: a backslash is always interpreted together with the character that immediately precedes it. Like other SAMPA diacritics, it is placed after the symbol to which it refers.

*In many computing applications the backslash is used as an escape character for the single **following** character.Ought we to follow this, and have a convention of backslash plus x, or stick to our SAMPA convention that diacritics follow the symbol they refer to, hence x plus backslash?*

Some solutions now immediately suggest themselves. For example, there are cases where the IPA uses a small capital corresponding to an upper-case character that SAMPA has preempted for another use. We define these with the backslash:

ʙ	small cap B, voiced bilabial trill (cf SAMPA [B] = IPA [β]	B\	92
ɢ	small cap G, voiced uvular plosive (cf SAMPA [G] = IPA [ɣ]	G\	
ʜ	small cap H, voiceless epiglottal fric. (cf SAMPA [H] = IPA [ʯ])	H\	
ʟ	small cap L, voiced velar lateral (cf SAMPA [L] = IPA [ʎ])	L\	
ɴ	small cap N, voiced uvular nasal (cf SAMPA [N] = IPA [ŋ])	N\	

One special case arises with the uvular r-sounds. SAMPA [R] is defined as covering both fricative/approximant, IPA [ʁ], and trill, IPA [ʀ]. Since [ʁ] is more widely distributed across languages than [ʀ], it seems more appropriate to keep SAMPA [R] for [ʁ], which leads to R\ for [ʀ]:

ʀ	small cap R, voiced uvular trill (cf SAMPA [R] = IPA [ʁ])	R\

Some further suggestions, for symbols from the main body of the IPA Chart:

ħ	crossed h, voiceless pharyngeal fricative, Arabic *ha*’	X\	
ʕ	reversed glottal stop, voiced pharyngeal fricative, Arabic *‘ayn*	?\	
ɦ	hooktop h, voiced glottal fricative	h\	
ʝ	turned f, barred dotless j, voiced palatal plosive, cf [J] = [ɲ]	J\	
ɮ	l-yogh ligature, voiced alveolar lateral fricative, cf [K] = [ɬ]	K\	
ɸ	phi, voiceless bilabial fricative	p\	
ʝ	curly-tail j, voiced palatal fricative	j\	
ɯ̽	turned m, right leg; velar approximant, cf [M] = [ɯ]		M\
ɹ	turned r, voiced (post-)alveolar approximant	r\	

From the last of these it follows that [ɻ], the voiced retroflex approximant, becomes SAMPA [r\`].

With these extensions, SAMPA is now able to deal with the phonemes of Arabic, the last major language previously uncatered for.

Various further symbols, from the lower part of the IPA Chart, can now be coded with the backslash in ways that will, I hope, be mnemonically helpful:

ʘ	bull's-eye, bilabial click	O\ *(letter O)*
ɘ	reversed e, close-mid central unrounded vowel, cf [@] = [ə]	@\
ɞ	closed epsilon, open-mid central rounded vowel, cf [3] = [ɜ]	3\

ɺ	turned long-leg r, voiced alveolar lateral flap	l\ *(letter l)*
ʢ	barred reversed glottal stop, voiced epiglottal fricative	<\
ʡ	barred glottal stop, epiglottal plosive	>\
ɕ	curly-tail c, voiceless alveolopalatal fricative	s\
ʑ	curly-tail z, voiced alveolopalatal fricative	z\
ʃ̡	hooktop heng, voiceless postalveolar and velar fricative	x\

The last three of these may be needed for general-phonetic or allophonic purposes only, since for the phonemic (or quasi-phonemic) notation of specific languages other solutions are preferred. In Polish, as suggested above, /ɕ, ʑ/ could be satisfactorily be represented as SAMPA /s', z'/ . For Japanese *sh* [ɕ], /S/ might well be acceptable. The best notation for the Chinese (Mandarin) *x q j* series awaits discussion, but /s' ts' dz'/ might be best. Certainly for Swedish *sj* [ʃ̡] the notation /S/ would be unambiguous.

We can even apply this solution for one or two symbols not currently recognized by the IPA:

ɪ	barred small cap I, central lax close unrounded vowel	I\
ʊ	barred upsilon, central lax close rounded vowel	U\

13. We have not yet found a way of coding most of the IPA **diacritics**. The best solution appears to be by choosing a character to bear the meaning "interpret the following character as a diacritic". Now SAMPA defines the asterisk (*) for special use as a **conjunctor**; but inquiry shows that this was not envisaged as appropriate for things like diacritics; rather, it ought to be reserved for special purposes that we may need in future applications as yet unformulated. Accordingly, for diacritics I now propose the use of the **underscore** (_ , ASCII/ANSI 95). Here are some obvious suggestions, shown as usual with another symbol as an example. Note that the symbol following the _ is assigned an interpretation different from its interpretation as a stand-alone SAMPA character.

t^{w}	raised w, labialized	t_w
$t^{ɣ}$	raised gamma, velarized, cf [G] = [ɣ]	t_G
$t^{ʕ}$	raised reversed glottal stop, cf [?\] = [ʕ]	t_?\
d^{n}	raised n, nasal release	d_n
d^{l}	raised l, lateral release	d_l

The suggested coding t_?\ preserves the convention that \ relates to the one character only that precedes it; and shows that in interpreting the operators \ and _ we must let \ take priority.

For **aspiration**, in phonemic transcription a simple h following the symbol for the consonant concerned will usually be adequate: thus in Korean we could represent the three plosive series straightforwardly as SAMPA /p, ph, pp/ (with or without /b/, strictly speaking allophonic, as required). In Hindi we can write /p, ph, b, bh/ as in the usual romanization. However, where a diacritic is explicitly needed, IPA [^{h}], we can adopt _h:

t^{h}	raised h, aspirated	t_h	42

14. Since we have proposed the apostrophe (ASCII/ANSI 39) as the sign for palatalization, we cannot use it in its traditional role as the **ejective** diacritic. Instead, we might use the visually similar > (ASCII/ANSI 62), preceded by an underscore:

pʼ	apostrophe, ejective	p_>	42 62

— and likewise t_>, k_>, s_> etc.

An alternative solution for the ejectives would be _?, thus p_? *etc. See below.*

This suggests a corresponding solution to the question of symbolizing the **implosives**, the sounds made with the other glottalic airstream. The hooktop in the IPA symbols ɓ, ɗ, ʄ, ɠ, ɠ is another pseudodiacritic comparable to the retroflex tail discussed above. I propose that we treat it likewise, using < (ASCII/ANSI 60) with a preceding underscore:

ɓ	hooktop b, implosive	b_<	42 60

and similarly d_<, J_<, g_<, G_< for ɗ, ʄ, ɠ, ɠ. If wanted, we can further write p_<, t_<, c_<, k_<, q_< for the voiceless implosives ɓ, ƭ, ƈ, ƙ, ʠ, from which IPA recognition was withdrawn in 1993.

15. Here are proposed codings for the **remaining diacritics** on the chart. As usual, they have been chosen wherever possible for their mnemonic appropriateness.

n̥	n_0 *(figure zero)*	under-ring, voiceless
s̬	s_v	subscript wedge, voiced
o̜	O_O *(capital O)*	subscript right half-ring, more rounded, cf O = ɔ
o̝	O_c	subscript left half-ring, less rounded
u̟	u_+	subscript plus, advanced
i̠	i_-	under-bar, retracted
ë	e_"	umlaut, centralized
ĕ	e_x	over-cross, mid-centralized
i̯	i_^	subscript arch, non-syllabic
b̤	b_t	subscript umlaut, breathy voiced
b̰	b_k	subscript tilde, creaky voiced
t̼	t_N	subscript seagull, linguolabial
t̪	t_d	subscript bridge, dental
t̺	t_a	inverted subscript bridge, apical
t̻	t_m	subscript square, laminal
d̚	d_}	corner, no audible release
d̴	d_e	superimposed tilde, velarized or pharyngealized

(but for ɬ we have already reserved SAMPA 5)

e̝	e_r	raising sign, raised
e̞	e_o	lowering sign, lowered
a̟	a_A	advancing sign, Advanced tongue root
a̠	a_q	retracting sign, retracted tongue root

The diacritics in the basic SAMPA set (= and ~, as in n=, A~ for IPA n=, A~) need no underscore; nor do the ' and ` proposed above. Since they are of frequent use it seems sensible to retain this keystroke-saving convention. However, no confusion would arise if on occasion it was wished to write them with an underscore, thus n_= etc. To avoid a possible source of error, none of them has been given any other definition in conjunction with _.

16. The underscore could in principle also be pressed into service to represent the IPA **tie bar**. The current chart mentions its use only for affricates and double articulations, and then only "if necessary". Some phoneticians also use it for diphthongs (for example, the author if the standard German pronunciation dictionary, Mangold 1990). The above diacritic codings have been chosen in such a way as to ensure there is no possible clash: _s, _S, _z, _Z, have been avoided, as have _p and _b and _i, _u, _y.

The qualification "if necessary" on the IPA Chart is important. Affricates can indeed normally be written in SAMPA as tS, dZ, ts, dz, etc; and if the extensions above are accepted then also ts\, dz\, (= IPA [tɕ, dʑ]) etc. If it is necessary to emphasize the status of affricate rather than sequence (cluster) of plosive plus fricative (as in the Polish *czy* vs. *trzy*), then it would theoretically be possible to write the affricate with the conjunctor, /t_S/, and the sequence without, /tS/; but in practice it would certainly be more satisfactory to assume affricate status unless an explicit **dis**junctor (hyphen) is used, thus /tS1/ *czy*, /t-S1/ *trzy*. The same applies with double articulations, where /kp/ and /gb/ do not normally need a tie bar, since the languages that use them do not admit the corresponding sequences. It also applies to diphthongs, which we may continue to write "aI, Au, 9y" etc. (or "aj, aw, aH" etc., depending on the phonology of the language in question), without any underscore.

17. Three of the five **click** symbols make use of standard characters within the basic range 33..127, namely [|, !, ||]. Since these symbols may be needed for prosodic or other purposes, it might be best to write them in SAMPA as |\, !\, ||\ respectively. Note that | and || involve the vertical 'pipe' symbol (ASCII/ANSI 124), not the diagonal 'slash' (ASCII/ANSI 47), which is reserved for other uses. The symbol |\|\ includes two successive | characters, which will not involve ambiguity, since no language has click clusters. The so-called palatoalveolar click , IPA [ǂ], could be written as =\. For the remaining click, the bilabial, IPA [ʘ], [O\] was suggested above.

18. **Alternative notations**. So many possible symbols become available if we adopt the \ and _ notations that we can even consider the luxury of permitting a few alternative symbolizations. Clearly, the adoption of SAMPA [5] for the frequently-needed dark-l symbol, IPA [ɫ], does not invalidate the alternative, analytic SAMPA notation [l_e]. The choice of [P] for the labiodental approximant, IPA [ʋ], is a response to colleagues' desire to have this useful symbol available directly from the keyboard, which led me to allocate it, rather than say [ɸ], to the spare shift-P

key; but this does not prevent us permitting the common-sense [v\] as an alternative if we wish. I suggested [_>] for the ejective diacritic so that parallelism could be preserved in the choice of [_<] for the implosive; but it would also alternatively be possible to write ejectives with [_?], in which case the link with glottal closure would be clearly signalled.

19. Three of the **suprasegmentals** on the chart fall within the basic SAMPA set: the primary and secondary stress marks " and % (= IPA " and ˌ), and the length mark : (= IPA ː). We can add

ˑ	:\	half-length mark
ĕ	e_X	breve, e<u>X</u>tra-short
‿	-\	linking mark

The minor and major group boundary symbols, | and ||, can remain as they are (pipe, ASCII 124).

20. Of the symbols on the 1993 IPA Chart, that leaves just the **tones and word accents** to be catered for. The SAM Prosody Group has already produced draft proposals for computer-readable transcription systems for the transcription of intonation, SAMSINT (Wells et al., 1992: 10-17). The combination of SAMSINT with the prosodic symbols of SAMPA yields the proposed SAM prosodic alphabet symbol set (SAMPROSA) (ibid.). However, there are various conflicts between the symbols of SAMPROSA and those of SAMPA. For example, as a segmental symbol, T and H are defined by SAMPA as a voiceless dental fricative and a labial-palatal approximant respectively, IPA [θ, ɥ]; but as a prosodic symbol, T is defined by SAMPROSA as denoting Top pitch and H as denoting High pitch. There are also conflicts with the IPA: SAMPROSA uses ! to show Downstep (as do many phoneticians), while in IPA it is a click symbol. None of this matters as long as prosodic and segmental transcriptions are kept apart from one another on separate tiers; but it does offer difficulties for a traditional linear phonetic transcription.

The IPA currently offers two sets of tone marks. One set involves accent marks written over the vowel letters, e.g. é (high tone), è (low tone). The other set involves a vertical tone bar to which a tone mark is attached at one of five height levels, e.g. ˥ (high tone), ˩ (low tone). In each case contour tones are symbolized by combining symbols for level tones, e.g. for low rising è᪲ or ˩˧. With a basic five levels of height (extra high, high, mid, low, extra low), each of these methods involves
a very large number of possible symbols, and the five contour marks shown on the Chart are merely a subset of those logically possible—there are ten possible rising contours (e.g. extra-low-to-mid, low-to-high), ten falls, and over a hundred possible three-level tones (e.g. a rise-fall, high-to-low-to-mid). Hence alongside the rising-falling ȇ or ˩˥˧ shown on the Chart, the "etc." that follows implies over two hundred other complex tone marks, of which a falling-rising ȇ᪲ or ˧˩˧, for example, is only one of the simplest.

Furthermore, many scholars still use the older IPA notation in which a fall was shown as [a᪲], a rise as [a᪲], etc. — the convention underlying both the conventional Pinyin transliteration of Chinese and the existing SAMPA recommendation to use grave and acute accent/apostrophe for these purposes.

Both the SAMPROSA and IPA systems imply a five-level analysis of pitch: top or extra high,

high, mid, low, and bottom or extra low. In SAMPROSA they are written respectively as T, H, M, L, B; in IPA as ˥́, ˦́, ˧̄, ˨̀, ˩̀ or as ˥, ˦, ˧, ˨, ˩.

21. It is not practicable to avoid overlap between symbols for segmentals (SAMPA) and for prosodics (e.g. SAMPROSA). We must therefore keep the two types of notation separate from one another by defining a **tier escape symbol** marking the point at which the notation moves from the segmental tier to the prosodic tier or back again. (More generally, we might wish to mix in other tiers: paralinguistic/discourse, morphological, etc.)

What should this tier escape symbol be? Perhaps the suggestion is the proposal by Gibbon (p.c.) to use paired angle brackets < > (ASCII/ANSI 60, 62), as often found in this sense in machine-readable corpus notations. (Since they might lead to ambiguities, angle brackets would not be suitable for material involving implosives, if the notation suggested in 14. above is adopted.)

Thus we would define

<	to mean	"start prosodic,paralinguistic, or other nonsegmental notation", and
>	to mean	"end prosodic etc. notation (return to segmental notation)"

The symbols could also be labelled if desired at each tier change, e.g.

or	<P	"start paralinguistic"
	<S	"start segmental" (or this might be the default and left implied by >).

Notwithstanding this, the basic SAMPA stress, length, and syllabicity marks [" % : =] can if desired be freely interspersed among segmental symbols without giving rise to any ambiguity. They do not necessarily require a tier escape symbol.

22. Thus using **unlabelled angle brackets** as a tier escape symbol into and out of prosodic transcription one might show the English word *nothing* with top pitch on the first syllable and low pitch on the second as

<T>"nVT<L>IN

where the angle brackets indicate that the first T is to be interpreted as "top pitch"; we then toggle back into segmental mode, signalling that the second T stands for a voiceless dental fricative [θ].

It remains to be seen whether colleagues wish to continue with SAMPROSA, or would prefer to adopt some version of the TOBI conventions.

Using SAMPROSA symbols and unlabelled angle brackets, the barline and arrow symbols on the IPA Chart are encoded as follows:

˥	<T>	Extra high, Top pitch
˦	<H>	High pitch
˧	<M>	Mid pitch
˨	<L>	Low pitch
˩	<B>	Extra low, Bottom pitch

ˊ	<BT>, *also* <LH>, *etc*	Rising
ˋ	<TB>, *also* <HL> *etc*	Falling
˥	<HT>	High rising
˩	<BL>	Low rising
ˆ	<HTH>, *also* <MHM> *etc*	Rising-falling
↓	<!>	Downstep
↑	<^>	Upstep
↗	</> *or* <R>	Global rise
↘	<\> *or* <F>	Global fall

23. Where it is wished to encode the IPA's alternative way of symbolizing tones and word accents, using accent marks over a vowel letter, the underscore could be used, together with SAMPROSA's B,L,M,H,T; R,F. These will not conflict with other proposals involving the conjunctor above. Thus:

e̋	e_T	Extra high
é	e_H	High
ē	e_M	Mid
è	e_L	Low
ȅ	e_B	Extra low
ě	e_L_H *or* e_R *or* e_/	Rising
ê	e_H_L *or* e_F *or* e_\	Falling
e̍	e_H_T	High rising
e̗	e_B_L	Low Rising
e̋̂	e_M_H_L *or* e_R_F *or* e_/_\	Rising-falling, etc.

24. It might be that in practice a **language-specific tone number** would sometimes be more convenient. This would mean that the four tones of Mandarin Chinese could be written

ma_1	*rather than as the more explicit*	ma_H	(high level, Pinyin *ma*),
ma_2	*rather than*	ma_M_H	(rise, *má*),
ma_3	*rather than*	ma_M_L_M	(fall-rise, *ma*),
ma_4	*rather than*	ma_H_L	(fall, *mà*);

or, equivalently, as <1>ma, <2>ma, etc., rather than as <H>ma, <MH>ma, etc.

By extending the SAMPROSA "nuclear tone" marks also to denote word or syllable tones, we could theoretically also write

tone 1	ma_-	*or* <->ma,
tone 2	ma_/ *or* ma_R	<'>ma *or* </>ma *or* <R>ma,
tone 3	ma__/ *or* ma_F_R	<`'>ma *or* <\/>ma *or* <FR>ma,
tone 4	ma_`	<\>ma *or* <F>ma.

—but this is an embarras de richesses. For Mandarin, I would vote for one of the first two

solutions. For Cantonese [fan], with its six contrastive tones, we could choose among

<1>fan	*or*	fa_1n	*or*	<TT>fan	(tone 1, high level)
<2>fan		fa_2n		<M>-fan	(tone 2, high rising)
<3>fan		fa_3n		<MM>fan	(tone 3, mid level)
<4>fan		fa_4n		<LB>fan	(tone 4, low falling)
<5>fan		fa_5n		<LM>fan	(tone 5, low rising)
<6>fan		fa_6n		<LL>fan	(tone 6, low level)

(Fortunately the mid front rounded vowel of Cantonese, which SAMPA transmutes into a numeral, is IPA [œ] = SAMPA [9], so outside the range of pitch/tone numbers.)

25. The Japanese **pitch accent**, phonetically a pitch downstep with contrastive function, is usually symbolized by Japanese scholars as ¬ (corner, ASCII 170, ANSI 172). To show downstep, SAMPROSA uses the exclamation mark, ! (ASCII/ANSI 33); to show upstep, ^ (ASCII/ANSI 94). Perhaps these would be acceptable as SAMPA symbols, too. They would not necessarily require a tier escape symbol. Thus *hana* 'flower' could be transcribed [ha^na!], but *hana* 'nose' as [ha^na]; *hashi* 'bridge' as [ha^Si!], *hashi* 'chopsticks' as [ha!Si], *hashi* 'end, edge' as [ha^Si] (although it is true that in Japanese the occurrence of the upstep is predictable, for which reason ^ could be omitted without ambiguity).

26. Taken as a whole, these proposed symbolizations, codings and conventions preserve the unique and useful SAMPA characteristic that symbol strings remain **uniquely parsable** even when written without spaces between successive characters.

Perhaps it is not out of place to make two general points about phonetic transcription. SAMPA, like the IPA in general, is merely a repertoire of symbols. Different phoneticians and researchers can make different use of them. In particular, the symbols are basically intended to symbolize **phonemes** (distinctive sound classes of a particular language) rather than to capture all the allophonic (subphonemic) detail that it is possible to observe, whether articulatory or acoustic. Thus in transcribing English phonemically we ignore the rather gross differences that exist in different phonetic environments: aspiration vs. nonaspiration of /p t k/, clearness vs. darkness of /l/, dental vs. alveolar vs. glottal place of /t/, and so on. It would be uneconomic to include all this allophonic detail in every piece of transcribed material, since it can be stated once and for all in a rules interpretation (allophonic rules). It is also possible to apply the IPA, or SAMPA, for other purposes such as allophonic transcription, the transcription of pathological speech, impressionistic transcription, and so on, but this is not the primary purpose. (In speaking of allophones, I am thinking of articulatory allophones: acoustic allophones are something else again.)

Equally, both SAMPA and the traditional IPA frequently use the same symbol for sounds in different languages, sounds that are objectively different from one another. When we write /t/ for English, its default realization is apicoalveolar and aspirated; when we write /t/ for French, its default is dental and unaspirated. In Swedish it is dental and aspirated; in Russian it is perhaps laminodental and velarized. It would be hopelessly uneconomic to insist on a different symbol for every "sound" that we can identify among the world's languages. It is better to use a simple straightforward transcription and to state the default values for a particular language in a general statement or rule, rather than to repeat this information over and over again, as we should be

doing if we insisted in making it appear throughout a running text. It is in this sense, for example, that the IPA uses the same symbol [ʃ] for Russian ш, Polish *sz*, English *sh*, and Italian *sci*, notwithstanding the very different resonances they have; and SAMPA should correspondingly use [S] for them all.

27. To finish, I attach a **tabulation** of all the proposed symbol codings.

no.	*SAMPA*	*IPA*	*SAMPA using *	*IPA*	*SAMPA using underscore*	*IPA*	
033	!	˥ Downstep	!\	!			
034	"	ˈ Pri. stress			a_"	ä	Centralized
035	#						
036	$						
037	%	ˌ Sec. stress					
038	&	Œ					
039	’	ʲ Palatalized					
040	(						
041	)						
042	* *general escape character ("conjunctor")*						
043	+				u_+	u̟	Advanced
044	,						
045	- *separator*		-\	‿	i_-	i̠	Retracted
046	.						
047	/ *indeterminacy in French*				a_/	ă	Rising
048	0				b_0	b̥	Voiceless
049	1	ɨ			a_1		*language-specific tone*
050	2	ø			a_2		" "
051	3	ɜ	3\	ɞ	a_3		" "
052	4	ɾ			a_4		" "
053	5	ɬ			a_5		" "
054	6	ɐ			a_6		" "
055	7	ɤ					
056	8	ɵ					
057	9	œ					
058	:	ː	:\	ˑ			
059	;						
060	< *tier escape*		<\	ʢ	b_<	ɓ	Implosive
061	n=	n̩ Syllabic	=\	ǂ			
062	> *tier escape*		>\	ʡ	t_>	t’	Ejective
063	?	ʔ	?\	ʕ	t_?\	tˤ	Pharyngealized
064	@	ə	@\	ɘ			
065	A	A			a_A	ą	Advanced TR
066	B	β	B\	ʙ	a_B	a̗	Extra low tone
067	C	ç					
068	D	ð					
069	E	ɛ					
070	F	ɱ			a_F	â	Falling
071	G	ɣ	G\	ɢ	a_G	aˠ	Velarized

072	H	ɥ	H\	ʜ	a_H	á	High tone
073	I	ɪ	I\	ɨ *(not IPA)*			
074	J	ɲ	J\	ʝ			
075	K	ɬ	K\	ɮ			
076	L	ʎ	L\	ʟ	a_L	à	Low tone
077	M	ɯ	M\	ɯɥ	a_M	ā	Mid tone
078	N	ŋ	N\	ɴ	t_N	t̪	Linguolabial
079	O	ɔ	O\	ʘ	O_O	o̹	More rounded
080	P	ʊ					
081	Q	ɒ					
082	R	ʁ	R\	ʀ	a_R	ă	Rising
083	S	ʃ					
084	T	θ			a_T	a̋	Extra high tone
085	U	ʊ	U\	ʉ *(not IPA)*			
086	V	ʌ					
087	W	ʍ					
088	X	χ	X\	ħ	e_X	e̋	Extra short
089	Y	Y					
090	Z	Z					
091	[						
092	\ *general diacritic*				a_\	â	Falling
093	]						
094	^	˖ Upstep			i_^	i̯	Non-syllabic
095	_ *introduces diacritic*						
096	@`	ɚ					
	d`	ɖ					
097	a	a			t_a	t̻	Apical
098	b	b					
099	c	c			O_c	o̜	Less rounded
100	d	d			t_d	t̪	Dental
101	e	e			d_e	đ	Velarized or pharyngealized
102	f	f					
103	g	ɡ					
104	h	h	h\	ɦ	p_h	p^{h}	Aspirated
105	i	i					
106	j	j	j\	ʝ			
107	k	k			a_k	a̰	Creaky voiced
108	l	l	l\	ɭ	d_l	d^{l}	Lateral release
109	m	m			t_m	t̼	Laminal
110	n	n			d_n	d^{n}	Nasal release
111	o	o			E_o	e̞	Lowered
112	p	p	p\	ɸ			
113	q	q			e_q	e̘	Retracted Tongue Root

114	r	r	r\	ɹ	E_r	ɛ̝	Raised
115	s	s	s\	ɕ			
116	t	t			b_t	b̤	Breathy voiced
117	u	u					
118	v	v	(v\	ʊ)	f_v	f̬	Voiced
119	w	w			t_w	tʷ	Labialized
120	x	x	x\	ʃj	a_x	ă	Mid-centralized
121	y	y					
122	z	z	z\	ʑ			
123	{	æ					
124	|	| *boundary*	|\	ǀ *click*			
125	}	ʉ			d_}	d̚	No audible release
126	A~	ɑ̃ Nasalized					

References

Esling, John H. and Gaylord, Harry, 1993. Computer codes for phonetic symbols. *Journal of the IPA* 23:2.

IPA, 1993. The International Phonetic Alphabet (revised to 1993). *Journal of the IPA* 23:1, centrefold.

Wells, J., Barry, W., Grice, M., Fourcin, A., and Gibbon, D., 1992. Standard Computer-Compatible Transcription. Esprit project 2589 (SAM), Doc. no. SAM-UCL-037. London: Phonetics and Linguistics Dept., UCL.

Revised draft 1995 04 28