This article was first published in *Computers in Genealogy* Vol. 6, No. 5 (March 1998)

Soundex - can it be improved?

by Peter Christian

The failure rate of Soundex

In an article in the September 1990 issue of *Computers in Genealogy* (Vol. 3, No. 7), Alan Stanier applied a number of statistical tests to two large sets of surname data in an attempt to establish the accuracy of Soundex coding.[1] The results were not encouraging. Of the more authentic test, run on the entire corpus of surnames in the 1851 census extract, Alan concluded that with the standard Soundex coding "only a third of the matches found will be correct, while a quarter of the correct ones will go undiscovered." Now the first of these shortcomings is, for genealogists at least, largely theoretical: for anyone prepared to go through thousands of entries in order to find a single match, a 33% success rate is actually very high. The real problem is the failure to locate 25% of genuine matches.

The obvious question is: can this failure rate be reduced? We can't, of course, expect to reduce it to zero. Surname pairs like CHOLMONDELY and CHUMLEY would only be matched as variants by a coding algorithm that ignores everything but the initial consonant, and even that would fail to match ASKEY and HASKEY, or LLOYD and FLOYD.[2] And of course, the shorter the code, the higher the number of incorrect matches.

Is Soundex sound?

There are really two parts to the question: is Soundex built on sound principles as far as it goes, and can it be easily extended to reduce the error rate?

---

[1] Alan Stanier, "How accurate is Soundex matching", *CiG* Vol. 3, No. 7, pp. 286-8. The text is available on the Web at *http://www.essex.ac.uk/AMS/articles/Soundex.html*

[2] In choosing examples for this article, I have relied on two main sources: P H Reaney, *A Dictionary of English Surnames*, 3rd edn, revised by R M Wilson, Oxford 1977, and a complete list of the surnames from the GRO marriage records for the March quarter of 1849, extracted from the transcriptions created by Mike Foster and others. The transcriptions can be downloaded from GENUKI (*http://www.cs.ncl.ac.uk/genuki/StCathsTranscriptions/#Complete1849MQ*), and I have put a file with the surnames, Soundex codes and frequency on the *CiG* Web site at *http://www.sog.org.uk/cig/vol6/1849mq.zip*. This is a comma separated file for importing into a database or spreadsheet.

The answer to the first question is in fact yes. Though the rules may at first sight seem arbitrary they are built on three main principles, each of which is perfectly sound:

1. **Some aspects of surnames are too variable to be reliably coded.**
   Vowels are the main area of dialect variation in English (much more so than the consonants), and it would be impossible to find a simple and reliable vowel-coding scheme that did not exclude obvious variants. The use of single or double consonants is highly variable, and largely not significant.

2. **Letters for similar sounding consonants are treated identically.**
   Many variants are distinguished by the replacement of one consonant by a similar one. For example <d> and <t> are interchangeable in some names, because they both represent, in phonetic terms, "dental stops".[3]

3. **Where a letter is used for different sounds, all other letters used for the same sounds must be coded identically.**
   This explains the apparent muddle of Group 2. Ideally this would be 3 distinct groups: one for s-like sounds, one for k/g-like sounds, and one for j/ch-like sounds. But because the letter <g> corresponds to both g and j sounds, and letter <c> corresponds to both s, k and ch sounds, there is no choice but to put all these into a single large group. <p, b> could perhaps be coded separately from <f, v>, but because <ph> can be used for <f> these two groups *have to* be combined.

Letter combinations

So far so good. But there are two fundamental principles of spelling that are completely ignored in Soundex, and which are particularly important in English.

1. **Some sounds are spelt not with single letters but with letter combinations (or with either).**
   For example, <dg> almost always represents a single sound, also spelt with <j>, and is not simply a combination of <d> and <g>. <ch> and <tch> spellings both correspond to single identical consonant sound, but are coded by Soundex as 2 and 32 respectively, making BACHELOR and BATCHELOR impossible to match with each other.

2. **Clusters of consonants are often not fully pronounced, and there is frequently a simplified spelling reflecting this.**
   For example, it is very common for 3-consonant combinations to be simplified: the [ndl] in LINDLEY will often sound like [nl] in casual pronunciation, and for that reason there is also a spelling LINLEY. Similarly WILTSHIRE ~ WILSHIRE, CHRISTMAS ~ CHRISMAS

Between them, these factors probably account for a significant number of Soundex's missed variants. And although the specific examples here are from English, the general principles hold for other European languages — it's just

---

[3] Letters in <angled brackets> represent *spelling*: letters in [square brackets] represent *sounds*..

<dg>	is coded by Soundex as 32, but almost always corresponds to the same sound as <g>, e.g. RODGERS~ROGERS (R326, R262), BURRIDGE~BURRAGE (B632, B620).
<ng>	usually corresponds to a single consonant in pronunciation, and should be treated as interchangeable with <n>, e.g. HAWKIN~HAWKING (H250, H252), CUMMIN~CUMMING (C550, C552).
<l>	is often not pronounced before another consonant (particularly after [a] or [o]): FOULKES~FOWKES (F422, F220), ROLF~ROFF (R410, R100), ALCOCK~AWCOCK (A422: A220), SHAWCROSS~SHALCROSS (S262: S426). A glance at Reaney & Wilson will show many surnames starting AL- which have AU- variants.
<r>	in many dialects of English, is not pronounced as a consonant before other consonants: for many speakers BORDEN and BAWDEN, though not related etymologically, sound identical (B635, B350). For many speakers there is no distinct [r] in the first syllable of CHARTERIS, (C636) so an r-less variant, CHATTERIS (C362), is not surprising.
<gh>	almost never corresponds to any pronounced consonant, thus LAUGHTON~LAWTON (L235, L350), LEIGH~LEE (L200, L000), BLIGH~BLY (B420, B400), and in many Irish names, e.g. SHAU(GH)NESSY.
<mb>	almost always represents simply an [m] in modern pronunciation, e.g. CROMB~CROOM (C651, C650), and the many names ending in COMB(E).
<nt, nd>	are often pronounced without the final <d, t> before a third consonant or at the end of a name: HANDSCOMBE~HANSCOMBE (H532, H525). Conversely, a <d> or <t> may be added to a name ending in [n], e.g. WYMAN~WEYMONT (W550, W552).
<m+Consonant>	are often separated by a so-called "glide" consonant <p, b>, as you can see if you say TOMKINS without making any special effort. Cf. HAMLIN~HAMBLIN (H545, H514). This is almost the reverse of the <mb> rule and leads to consistent variation in Soundex between names with ..5.. and ..51.. codes.

Table 1: Problem consonant combinations

that in English the situation is exacerbated by some highly archaic spelling conventions, particularly in names.

Table 1 gives some specific examples of Soundex's failure to cope with consonant combinations. This is not, as it may seem at first sight, just a list of oddities — they are regular rules for the correspondence between sound and spelling in English, and are well documented in studies of English orthography.[4] But because they relate to a consonant in its context and not in isolation, the relationship between the surnames affected by these rules is completely invisible to Soundex.

---

[4] For example, Edward Carney, *A Survey of English Spelling*, London 1994. E J Dobson, *English Pronunciation 1500-1700*, 2 vols, 2nd edn, Oxford 1968.

To get some idea of the effect of the potential contribution of these conventions to the failure rate of Soundex, I looked at a sample of English surnames, namely the list of surnames in the March 1849 GRO marriage indexes, to see how frequent the letter combinations were (see Table 2). This does not, of course, indicate the actual failure rate of Soundex for these combinations, but it does give an idea of the number of names in which either (a) there might be a regular alternative spelling, or (b) there is the potential for a clergyman or a census enumerator to use a spelling which would give a different Soundex code.

Table 2: Frequency of consonant combinations

Spelling	alternative pronunciation or spelling	normally coded as	suggested recoding	% in 1849 sample
ng	n	52	5	4.9%
gh	*zero*	2	0	2.0%
mp, mb	m	51	5	2.0%
nt, nd	n	52	5	5.8%
tch/dg	c/g	32	2	2.4%
l+cons.	w+cons. *zero*+cons.	4	0	8.9%
Surnames with one or more of these combinations				**18.8%**

Of course, some of these occur too late in a name to affect a Soundex code, but the fact that almost a fifth of names in the sample have one or more of these combinations suggests that these might explain a significant portion of the 25% failure rate found by Alan Stanier. And of course there are other combinations which regularly produce variants, such as those mentioned above.

Final <s>

There is a single problem consonant at the end of surnames, which I suspect contributes very highly to the failure rate. Many English surnames have variants with and without a final <s>.[5] This is particularly the case with surnames derived from male forenames, e.g. JOHN~JOHNS (J500, J520), but it is also common with other names, e.g. HILL~HILLS (H400, H420) or DOWN~DOWNES (D500, D520). With about 8.5% of names in the 1849 sample ending in consonant+<s> or consonant+<es>, this must account for some of the overlooked matches. Although, because the <s> is at the end of the name it is sometimes irrelevant to Soundex, coming after the last

[5] There is a good discussion of these names in the Introduction to Reaney & Wilson, pp. xxxiii ff.

significant letter (e.g. ATKIN~ATKINS, both A325), even a superficial examination of the list from the 1849 sample shows that most of these names are fairly short.

Intials

Finally, there is a problem with the way that Soundex codes the start of names: <h> and vowels are normally ignored in Soundex coding, and yet they are treated like any consonant at the beginning of a name. But <h> is often dropped in English speech, which gives rise both to stable h-less variants, and to misunderstandings by recorders of oral information (e.g. ASKEY~HASKEY A200, H200). And since one of the reasons for ignoring vowels in the first place is their large variability, it is rather inconsistent for Soundex to treat this as unproblematic at the position in the word where it matters most for searching. While it may be the case that there is less vowel variation initially than later in the word, it is not hard to find pairs like AMERY~EMERY (A560, E560) or AYERS~EYRES (A620, E620).

Of course, when these sorts of variation combine, Soundex completely breaks down: ENNION, INIONS, ANYON, ONIONS, HENNION are all given under ENNION in Reaney & Wilson, but each has its own Soundex code!

Possible improvements

With an understanding of why Soundex fails to match variants when it does, we are now in a position to see if it can be improved, or whether the only hope of improvement is a coding scheme based on quite different principles.

The problems at the beginning and end of names could be overcome simply by adding further “single letter” rules to Soundex, e.g.

- initial <h> should be ignored;
- all initial vowels (including those after an ignored <h>) should be coded with a single code — I suggest “A”;
- final <s> should be dropped (possibly just after a consonant, or consonant+<e>).

But the problems arising from consonant combinations in the middle of names are more complex. Some of them could be dealt with by having some context-sensitive “delete” rules. Just as Soundex already deletes the second of two consonants with the same code, it could delete <b, p> after <m>, for example.[6] However, it would have to do so *before* the operation of the

[6] The problem is that Soundex uses only a single dimension of similarity, whereas English consonants have three (type, position, and voicing). There is a case for assigning <m> and <n> to

existing Soundex rules: we may want to delete <d> before <g>, but we certainly do not want to delete every Group 3 letter before any Group 2 letter. Similarly, if <gh> is going to be deleted, this will need to be done before the <h> itself is removed.

However, it would be preferable to have more flexible handling of letter combinations: in getting rid of the <d> in EDGE, it would be nice not to lose the <d> in LYDGATE, for example.

Pronunciation

This leads us to a final observation. With the exception of the final <s> rule, the types of variation I have looked at have something important in common: they arise from having *some* spellings which are closer to actual pronunciation than others. Indeed, the spellings more remote from pronunciation are typically an archaic record of earlier pronunciation — one of the reasons for the amount of variation in English surname spellings is that they have escaped the general standardisation of spelling which began in the 17th Century.

The fact is that people whose surnames are spelt THOMSON and THOMPSON mostly pronounce their names identically, and those called CHOLMONDELY and CHUMLEY, or LAUGHTON and LAWTON certainly do.[7] With knowledge of how the sounds and spellings of English have developed, it should not be impossible to devise a coding scheme which reduces all names to a form closer to their current pronunciation before putting them through a Soundex-like algorithm which removes remaining irrelevant distinctions.[8]

At what price?

Of course, greater success in matching variants comes at a price. There are two disadvantages to what I am suggesting. The first is that removing distinctions to improve variant matching, unless it is exceptionally accurate, will also reduce distinctions between names that are *not* variants, with a

---

groups 1 and 2 respectively (i.e. using a different dimension), which would make these rules unnecessary.

[7] Where obviously archaic spellings match current pronunciation, it is usually a case of what is called "spelling pronunciation", i.e. literate folk have started to use the spelt form as the basis for pronunciation rather than vice versa. The 'z' pronunciations of names like MENZIES and DALZELL are extreme examples – the 'z' is a printer's "mistake" for the archaic letter yogh, which stood for a sound like 'y'. There is not and never has been a linguistic basis for a 'z' pronunciation of these names.

[8] Carney, Chapter 4, gives text-to-speech rules for every letter, which could form the basis of such a program.

corresponding increase in incorrect matches. However, the effort required to overlook or delete these is as nothing compared to the inconvenience of failing to find matching names. And some of the incorrect matches derive from the over-inclusive nature of Group 2. With a pronunciation-based system, Group 2 could be split into two or three separate groups, with a consequent reduction in incorrect matches.

The second problem is that while some of the suggested rules are equally applicable to other languages and orthographical traditions, some are specific to English surnames and records. For example, the <dg> and <gh> rules are results of the specific history of English, while the rules about three-consonant combinations reflect much more general tendencies in language. Of course, the spelling traditions of other languages may have counterparts to our <gh> rule — a system for French would have other spellings which are not reflected in pronunciation. In particular, one would need to look at the effect of such rules when applied to US records, which include names from a wide variety of spelling traditions. However, Soundex is already heavily biased towards English, and has a number of faults when applied to other languages, so it cannot usefully function as a universal surname encoder.

Other benefits

While I have concentrated here on stable surname variants, it is worth noting that a significant proportion of the "errors" that enter records when they are created from oral testimony can be explained by the features I have discussed.[9] This is particularly the case for times, before universal education, when most respondents would have been unable to reply to the question, "How do you spell that?", and the idea that a surname like BURRAGE is being mis-spelt if it is recorded as BURRIDGE, or LINDLEY recorded as LINLEY, would have been simply anachronistic.

If we had a coding system which, before ironing out the type of variation that Soundex already copes with, reduced surnames to their approximate pronunciation, there would be less variation in the input to the Soundex algorithm, and Soundex's own inability to deal with these forms of variation would be unimportant.

[9] Errors in the copying of written records, however, are quite different: errors are introduced mainly through the omission or transposition of letters, and the misreading of the handwriting.

Conclusion

What I have not been able to do here is provide a thorough statistical analysis of the extent to which these suggestions would provide an improved surname coding system, nor have I made any attempt to gauge how complex it would be to implement such a system. However, I hope I have shown that improvements to Soundex *are* possible in principle, not least because a proportion of its failures can almost certainly be ascribed to a group of general failings, each of which is capable of being formulated in a programmable rule.

Further information

The origins of Soundex are discussed in Alan Stanier's article and in subsequent letters in Vol. 3 of *CiG*, from Leonard H. Smith Jr on p. 342 and from Richard L. Halliday on pp. 430-32. There is also a letter from Michael Atyeo on p. 392 on Phonebase and Soundex.

A number of Soundex programs have been published in earlier issues of this magazine, but all are for dialects of Basic which are no longer available on current computers.

In addition to Barney Tyrwhitt-Drake's Soundex utility described in this issue, a number of other Soundex programs are available, and there are on-line Soundex converters at:

http://searches.rootsweb.com/cgi-bin/Genea/soundex.sh

http://www.geocities.com/Heartland/Hills/3916/soundex.html

The NARA (National Archives and Record Administration) has a Web page on Soundex:

http://www.familyhistory.com/faqs/narasdex.htm