11
Phonetic Coding Systems for Names

Phonetic coding systems use the way words or syllables are pronounced when spoken to help reduce minor typographical errors. Soundex and the New York State Identification and Intelligence System (NYSIIS) are two widely used phonetic schemes for encoding names. NYSIIS results in substantially more codes than does Soundex, and is harder to describe. Although NYSIIS provides many more codes, individual NYSIIS and Soundex codes associated with commonly occurring surnames such as Smith, Johnson, Brown, and Martin have approximately the same number of records associated with them. Neither Soundex nor NYSIIS can deal with most insertions, deletions, or transpositions of consonants. The primary value of both Soundex and NYSIIS in record linkage is to assist in bringing together records – that is, in *blocking* records. In fact, Jaro [1989] suggests using the Soundex version of names of individuals only as a blocking variable. He argues that using it for other purposes can create problems because many different names have the same Soundex code. In Chapter 12, we provide a detailed description of blocking and include several examples.[1]

11.1. Soundex System of Names

Soundex is an algorithm devised to code people's last names phonetically by reducing them to the first letter and up to three digits, where each digit represents one of six consonant sounds. This facilitates the matching of words (e.g., names of individuals or names of streets) by eliminating variations in spelling or typographical errors. Such schemes might be usefully employed with airline reservations systems or other applications involving people's names when there is a good chance that the name will be misspelled due to poor handwriting or voice transmission. For example, Soundex would treat "Smith," "Smithe," and "Smyth" as the same name. The standard Soundex algorithm works best on European last names. Variants have been devised for names from other continents/cultures/ethnic backgrounds.

[1] The string comparator metrics discussed in Chapter 13 are much better than coding schemes at comparing two strings that are brought together. However, string comparator metrics are of little use in blocking records together (except with relatively small files).

11.1.1. *History of Soundex*

Beginning in 1935, the Social Security Administration offered old-age pensions to Americans who could show they were at least 65 years of age. Unfortunately, many of the seniors eligible for such pensions had neither birth certificates nor any other means of establishing their date of birth. For example, a senior reaching her 65th birthday during 1935 would have been born in 1870. Because there were few statewide or even countywide birth registration systems in operation in 1870, the Social Security Administration asked the US Congress to help provide an alternative method of establishing a person's age from official records.

The solution was to obtain the requisite birth dates from US Decennial Censuses. The job was assigned to the Works Progress Administration (WPA). However, the Federal government's first need was for an index to enable it to get easy access to the records of the Decennial Censuses. The system selected to index names from the censuses was called "Soundex" – a coding system that codes names according to their sound.

11.1.2. *History of Census Bureau Soundex Indices*

According to Dollarhide,[2] The Soundex indexing system "was used to create heads of household name indexes to the 1880, 1900, 1910, 1920, and 1930 [Decennial] Censuses."

"The 1880 Soundex index was compiled for all states but lists only families in which a child of 10 years or younger was included. The 1900 and 1920 indexes are complete indexes to all heads of household for all states. The 1930 Soundex was compiled only for twelve southern states only."

"In the 1910 census index, only 21 states were indexed. Of these, 15 of the states were indexed under the name 'Miracode,' and 5 states were done using the Soundex name. The coding of last names in Soundex and Miracode was identical. In fact, the only difference between Soundex and Miracode was in the citation of a family's position on a census page."

"All of the Soundex indexes were originally hand-written on $3'' \times 5''$ index cards, each card showing a head of household by full name and a list of all other persons residing in a household. Persons with a different surname than the head of household usually had a separate index card, coded under their own surname, as well as the one for the household in which they were listed. Included was each person's name, age, relationship to the head of household, nativity [i.e., place of birth], and a reference to the location in the census schedules where that family appears."

[2] The material cited in this section was taken from some Internet sites that no longer exist. The author of that material is Dollarhide who is also the author of *The Census Book* that is available on the Internet at www.heritagequestonline.com/prod/genealogy/html/help/census_book.html

inistration offered old-age pensions
at least 65 years of age. Unfortu-
ı pensions had neither birth certifi-
their date of birth. For example, a
35 would have been born in 1870.
ıuntywide birth registration systems
ıinistration asked the US Congress
ɔlishing a person's age from official

e birth dates from US Decennial
cs Progress Administration (WPA).
ɔd was for an index to enable it to
ıial Censuses. The system selected
l "Soundex" – a coding system that

ı Soundex Indices

lexing system "was used to create
880, 1900, 1910, 1920, and 1930

ɔr all states but lists only families in
ɔluded. The 1900 and 1920 indexes
ld for all states. The 1930 Soundex
ɜ only."
were indexed. Of these, 15 of the
ɔde,' and 5 states were done using
ıes in Soundex and Miracode was
ı Soundex and Miracode was in the
ıge."
ıally hand-written on $3'' \times 5''$ index
l by full name and a list of all other
ı a different surname than the head
ɔd, coded under their own surname,
ch they were listed. Included was
ɔ head of household, nativity [i.e.,
ion in the census schedules where

n from some Internet sites that no
ollarhide who is also the author of
ɪnet at www.heritagequestonline.com/

"The Works Progress Administration (WPA) compiled the 1880, 1900, 1920, and 1930 census indexes in the late 1930s, while the Age Search group[3] of the Census Bureau compiled the 1910 Soundex/Miracode index in 1962. Only the 1910 Miracode names were entered into computers for sorting purposes. This was done by keypunch in 1962, and the data [were processed by mainframe computers, with the output consisting of a printed strip] for each entry. The microfilmed images of the 1880, 1900, 1920 and 1930 Soundex are handwritten cards, while the images of the 1910 Miracode are from computer-generated printed strips. (For the five 1910 Soundex states, the originals were the same type of $3'' \times 5''$ index card as the other Soundex indexes; and the bottom of each printed card has the year 1962 on it.)"

Every Soundex index is now available to the public.

11.1.3. Soundex Indexes

"In all Soundex indices, the original cards were sorted by the Soundex code for the head of household's surname, then alphabetized by each person's first name. As a result, all [surnames] with the same Soundex codes are interfiled." For example, all last names with the code L000 (e.g., Lee, Leigh, Low, Law, Liem, and Lieh) are interfiled. So if you know the Soundex code and the first name of an individual, you can go directly to that individual's index card.

11.1.4. Soundex Coding Rules

The algorithm below implements a modern version of the Soundex coding system, a technique that Knuth [1998] attributes to Margaret K. Odell and Robert C. Russell [see *US Patents 1261167* (1918) and *1435663* (1922)].

SOUNDEX ALGORITHM

Step1: The first letter of a last name retains its alphabetic designation.
Step 2: In other than the first position, the letters a, e, i, o, u, y, w, and h are not coded at all. (This eliminates the vowels and the mostly silent sounding letters from the coding scheme.)
Step 3: The remaining consonants – the hard consonants – are coded according to the table below.

[3] During that period the Census Bureau set up a special "Age Search" group tasked with using "the Soundex Indexes to locate a person in one of the censuses. The Age Search group researched applications for Social Security pensions from seniors who could not prove their age. Upon finding someone in the census, the group would record the person's name, [place of birth], and age information and then issue a substitute for a birth certificate." "The Age Search group still exists, and its primary function remains" to enable individuals to verify their age via information obtained from prior censuses.

Coding Guide

Code	Key letters and equivalents
1	b, p, f, v
2	c, s, k, g, j, q, x, z
3	d, t
4	l
5	m, n
6	r

Step 4: If two or more letters with the same code were adjacent in the original name (prior to Step 1), or adjacent except for intervening h's and w's, then only the first letter is coded.
Step 5: Every Soundex code must consist of one alphabetic character followed by three digits. So, if there are less than three digits, an appropriate number of zeroes must be added; if there are more than three digits, then the appropriate number of the rightmost digits must be dropped.

Example 11.1:

To illustrate Steps 1, 2, and 5, we consider names such as Lee and Shaw. A name consisting of all vowels after the first letter, such as Lee, would be coded as L000 because the first letter, "L", is always retained, the two vowels, both "e", are dropped, and three zeroes are appended to attain the required three digits. Similarly, Shaw would be coded as S000 because the first letter, "S", is retained, the vowel, "a", is dropped as are the "mostly silent sounding letters" "h" and "w", and three zeroes are again appended to attain the required three digits.

Example 11.2:

To illustrate Step 4, we consider names such as Gauss, Cherry, and Checker. Gauss is coded as G200 because the two vowels are dropped, the double "s" is treated as a single letter, and two zeroes are appended to attain the requisite three digits. Similarly, Cherry is coded as C600 because the two vowels and the "h" are dropped, the double "r" is treated as a single letter, and two zeroes are appended to attain the requisite three digits. Finally, Checker is coded as C260 because the two "e"s and the "h" are dropped, the pair of letters "ck" is treated as a single letter since they have the same code (namely, "2"), the "r" is coded as a "6," and one zero is appended to attain the requisite three digits.

Example 11.3:

For a more complex example, we consider a name such as "Coussacsk." Here, the three vowels are dropped, the double-letter sequence "ss" is coded as "2" and the three-letter sequence "csk" is also coded as "2" because these three consecutive letters all have the same code of "2." So, "Coussacsk" is coded as C220.

Names

letters and equivalents
, f, v
k, g, j, q, x, z

n

ıe same code were adjacent in the original
ccept for intervening h's and w's, then only

onsist of one alphabetic character followed
than three digits, an appropriate number of
nore than three digits, then the appropriate
be dropped.

nsider names such as Lee and Shaw. A name
irst letter, such as Lee, would be coded as
s always retained, the two vowels, both "e",
ppended to attain the required three digits.
S000 because the first letter, "S", is retained,
he "mostly silent sounding letters" "h" and
ɔended to attain the required three digits.

names such as Gauss, Cherry, and Checker.
the two vowels are dropped, the double "s"
ro zeroes are appended to attain the requisite
ɔded as C600 because the two vowels and the
treated as a single letter, and two zeroes are
ree digits. Finally, Checker is coded as C260
are dropped, the pair of letters "ck" is treated
he same code (namely, "2"), the "r" is coded
d to attain the requisite three digits.

consider a name such as "Coussacsk." Here,
e double-letter sequence "ss" is coded as "2"
sk" is also coded as "2" because these three
same code of "2." So, "Coussacsk" is coded

11.1.5. Anomalies with Soundex

Some names that are closely related are coded differently. For example, while "Lee" is coded as L000, "Leigh" is coded as L200; "Rogers" is coded as R262 while "Rodgers" is coded as R326; and "Tchebysheff" is coded as T212 while "Chebyshev" is coded as C121.

On the other hand, some unrelated names have identical codes in Soundex. For example, both "Lee" and "Liu" are coded as L000; both "Gauss" and "Ghosh" are coded as G200; and both "Wachs" and "Waugh" are coded as W200.

Another issue is how to treat names such as Lloyd, van Buren, or von Munching.

Despite the problems noted above, Knuth [1998] concludes that "by and large the Soundex code greatly increases the chance of finding a name in one of its disguises."

11.2. New York State Identification and Intelligence System (NYSIIS) Phonetic Decoder

In this section, we present a seven-step procedure for implementing the NYSIIS coding scheme as Taft [1970] originally proposed it. In the first step, the initial letter(s) of a surname are examined and altered as necessary. In the second step, the same is done for the last letter(s) of the surname. In Step 3, the first letter of the NYSIIS coded surname is established. Steps 5 and 6 constitute an iterative procedure for creating the remaining letters of the NYSIIS-coded surname. In this iterative scheme, we begin with the second letter of the altered surname and scan each letter of the remaining letters of the surname using an imaginary "pointer." In Step 5, one or more of the letters of the coded surname are established via a set of "rules." The rules are reapplied each time the "pointer" is moved to the next letter of the name. In Step 7, the end portion of the NYSIIS code just created is subjected to a further check and changed as necessary.

In an in-depth comparative study of five coding systems, Lynch and Arends [1977] judged a "modified" NYSIIS coding scheme to be the best surname coding system because:

> It (1) placed variations of a given surname in the same code, (2) limited the size of each code, and (3) created codes that contain few dissimilar names.

The individual steps of the NYSIIS coding scheme are as follows, with coding beginning in Step 3:

Step 1: Change the initial letter(s) of the surname as indicated in the table below:

Changing the initial letter(s) of the surname

Original letter(s)	Altered letter(s)
MAC	MCC
KN	NN
K	C
PH	FF
PF	FF
SCH	SSS

Step 2: Change the last letter(s) of the surname as indicated in the table below.

Original letter(s)	Altered letter
EE	Y
IE	Y
DT	D
RT	D
RD	D
NT	D
ND	D

Step 3: The first character of the NYSIIS-coded surname is the first letter of the (possibly altered) surname.
Step 4: Position the "pointer" at the second letter of the (possibly altered) surname.
Step 5: (Change the current letter(s) of the surname – i.e., the one at the present position of the "pointer".) Execute exactly one of the following operations, proceeding from top to bottom:

(a) If blank, go to Step 7.
(b) If the current letter is "E" and the next letter is "V," then change "EV" to "AF."
(c) Change a vowel ("AEIOU") to "A."
(d) Change "Q" to "G."
(e) Change "Z" to "S."
(f) Change "M" to "N."
(g) If the current letter is the letter "K," then change "K" to "C" unless the next letter is "N." If "K" is followed by "N," then replace "KN" by "N."
(h) Change "SCH" to "SSS."
(i) Change "PH" to "FF."
(j) If "H" is preceded by or followed by a letter that is *not* a vowel (AEIOU), then replace the current letter in the surname by the preceding letter.

(k) If "W" is preceded by a vowel, then replace the current letter in the surname with the preceding letter.

Step 6: The next character of the NYSIIS code is the current position letter in the surname after completing Step 5 (but omitting a letter that is equal to the last character already placed in the code).

After putting a character into the code, move the pointer forward to the next letter of the surname. Then return to Step 5.

Step 7: (Change the last character(s) of the NYSIIS-coded surname.) If the last two characters of the NYSIIS-coded surname are "AY," then replace "AY" by "Y." If the last character of the NYSIIS-coded surname is either "S" or "A," then delete it.

Example 11.4:

We illustrate the use of the NYSIIS coding system in the examples summarized in the following table:

Examples of use of NYSIIS coding system[4]

Surname	NYSIIS-coded surname
Brian, Brown, Brun	Bran
Capp, Cope, Copp. Kipp	Cap
Dane, Dean, Dent, Dionne	Dan
Smith, Schmit, Schmidt	Snat
Trueman, Truman	Tranan

11.3. Where Are We Now?

In this chapter we discussed two coding schemes that can be used to enhance record linkages. These schemes have primary application in blocking – the subject of our next chapter.

[4] All of these examples are taken from Newcombe [1988].