Phonetic Matching Toolkit with State-of-the-art Meta-Soundex Algorithm for English and Spanish Words

Keerthi Koneru, Cihan Varol

 Department of Computer Science

Sam Houston State University, Huntsville, TX 77341, USA

Contact: keerthi@shsu.edu, cvarol@shsu.edu

*Abstract*—Researchers confront major problems while searching for various kinds of data in the large imprecise database, as they are not spelled correctly or in the way they were expected to be spelled. As a result, they cannot find the word they sought. Over the years of struggle, pronunciation of words was considered to be one of the practices to solve the problem effectively. The technique used to acquire words based on sounds is known as “Phonetic Matching”. Soundex is the first algorithm proposed and other algorithms like Metaphone, Caverphone, DMetaphone, Phonex etc., are also used for information retrieval in different environments. This project mainly deals with the analysis and implementation of newly proposed MetaSoundex algorithm for English and Spanish languages which retrieves suggestions for the misspelled words. A phonetic matching toolkit will also be developed enclosing the different phonetic matching algorithms for both Spanish and English languages.

*Index Terms* — Caverphone, DMetaphone, Information retrieval, Misspelled words, Metaphone, MetaSoundex, NYSIIS, Phonetic matching, Soundex, Spanish Soundex, Spanish Metaphone.

# I. Motivation

Information deterioration is an intensive problem for every organization in the present era. With the increase in the amount of information saved day by day, there is desperate need for locating the mistyped data. Organizations are facing great challenge to maintain the quality of data in information systems with various sources of data damage. Whenever the data is assimilated from multiple sources, it is complicate to recognize the duplicate records due to the existence of misspelled data for the same record. As a result, the information of organization always ends up at risk.

Databases play a crucial role in almost all the establishments. Matching records in database is a persistent and well-known problem for years. One of the techniques to improve the data quality, which uses variations of sound to detect the misspelled data, is Phonetic matching. The evolution of phonetic matching has come into frame when there is a hardship in the retrieval of information. The technique of obtaining words using sounds was used in the US census since the late 1890’s, but a concrete solution to this was first proposed and patented by Robert C. Russell in 1912 as Soundex algorithm for the English language. Later, many algorithms were developed based on the different specifications and language constraints. Some of the other prominently used algorithms are Metaphone, DMetaphone, Caverphone, and New York State Identification and Intelligence System (NYSIIS) Phonetic code. Phonetic matching plays a key role in information retrieval in multilingual environments, where diversities in pronunciation or writing styles with same meaning may be present. In such cases, the phonetic matching is also used for different languages other than English. Though data is available in many languages, according to the recent census of 2014, Spanish and English occupy the second and third places, respectively, among the most widely spoken languages in the world [1].

Though there are several algorithms, no algorithm is concrete as each and every algorithm has its own disadvantages [2]. This paper involves proposal and implementation of a new hybrid algorithm, namely MetaSoundex, which overcomes the main disadvantages of Soundex and Metaphone. It also encompasses the development of a web application tool which can provide approximate matching for the misspelled words of English and Spanish languages using different phonetic matching algorithms, specifically Soundex, Metaphone, MetaSoundex implemented in both English and Spanish Languages. Other phonetic algorithms such as DMetaphone, Caverphone, and NYSIIS will also be implemented and analyzed for English language. The efficiency of these algorithms is calculated by obtaining information retrieval metrics – Precision and Recall.

The rest of this paper is structured as follows. Related works are described in next section. Section III introduces phonetic matching techniques used in this study and the architectural implementation of proposed algorithm. The Timeline for the project is presented in section IV.

# II. Background

Data matching process mainly involves comparison of records to ascertain whether they are same entity or not. Phonetic comparison meticulously obtains the quantitative analysis of pronunciations between speech forms and spellings of words. The different sources of variations can be illustrated as:

(1) Spelling variations

(2) Phonetic variations

(3) Double names or double first names

(4) Change of name [2].

Due to different criteria mentioned above, rather than looking for exact matching, approximate matching would be worthwhile.

The early algorithm is Soundex, which produces a four digit code retaining its first letter. It is used as a standard feature in applications like mySQL, oracle, etc.

Because of the few disadvantages like dependency on the first letter, failure of detection of silent consonants, Soundex can only be used in applications where high false positives and false negatives can be tolerated [2].

An improvement of Soundex is implemented by Beider and Morse to reduce the number of false positives and false negatives, known as Beider-Morse Phonetic Matching (BMPM). Beider et.al. [3], has also mentioned that the algorithm is extended to languages other than English, with the application of some generic rules to obtain the phonetic codes. Varol et.al. [4], discussed BMPM as a hybrid technique with a 6-letter encoded code in which the percentage of irrelevant matches can be abated by 70%.

Phonex is a technique in which words are pre-processed before encoding. In order to overcome defects of Phonex, Phonix has been introduced with a number of transformations in the beginning, ending and in the middle of the word [4].

NYSIIS algorithm was developed in 1970 as a part of New York State Identification and Intelligence System project headed by Robert L. Taft [5]. The algorithm produces a canonical code similar to Soundex, but generates only alphabetic code [6]. Balabantaray et.al mentioned that unlike Soundex, NYSIIS retains information regarding vowels [6].

Though sounds are taken into consideration, all the above mentioned algorithms consider the phonetics of each letter. A new technique which considers diphthongs (combination of two or more letters) of words was first developed by Lawrence Philips in 1990 known as Metaphone. Bhattacharjee et.al [7] has stated that the technique is mainly used for data cleaning in the text files to remove erroneous data.

Pande et.al [8] detailed that the Metaphone has its extended usage in stemming, which improves performance in Information Retrieval (IR). In stemming, natural language processing tools like Levenshtein Edit Distance (LED) algorithm conflated with phonetic matching algorithms like Metaphone are used for greater accuracy [8]. David Hood cited that though the algorithm is sensitive to combination of letters like ‘TH’, it is not subtle enough with the vowels especially at the postvocalic L and R [5].

Double Metaphone, popularly known as DMetaphone, is an enhancement to Metaphone algorithm by Lawrence Phillips in 2000. Unlike Soundex, which encodes letter by letter, DMetaphone encodes groups of letters called diphthongs according to a set of rules [4]. Carstensen et.al, mentioned that the algorithm is more effective while matching proper names and short sentences in databases [9].

In pace, the specified algorithms are not suitable for a particular database, named Caversham, which is mainly used for data source linkage. The algorithm, known as Caverphone, which is analogous to Metaphone with some rules subsequently applied, is enforced by David Hood in 2002 to encode the data of Caversham database [5]. The algorithm is later improvised in 2004 to Caverphone 2.0, to increase its accuracy and efficiency by applying more set of rules. David Hood [5] also stated that the algorithm is efficient by giving precise matches when compared to Soundex and Metaphone algorithms for linking data sources.

One of the major applications of phonetic matching algorithms is its appliance to different languages. The limitations of Soundex make it straightforward that the algorithm is specifically designed for English language. Also, the grouping articulation of the English letters and limit to the four characters makes it less efficient to detect common spelling errors in other languages such as Spanish [10]. Hence in 2012, Am´on et.al [11] have proposed an improvement to Soundex algorithm by including Spanish letters making it feasible to obtain phonetic codes for Spanish words. In the same year, Alejandro Mosquera has developed Metaphone algorithm for Spanish language by adapting the techniques from the algorithm used for English Language [12].

Advanced techniques with collaboration of two or three algorithms have paved the way for obtaining codes to the phonetics of a word in different languages [4]. Shah et.al [2] presented an overview of different algorithms and mentioned a comparative study on phonetic matching algorithms. They mainly focused on the brief description of issues in Soundex algorithm and concluded that each of the discussed algorithms has its own disadvantages in particular areas.

Soundex and Q-Gram can be applied on Indic languages like Hindi, Marathi etc. In order to increase the efficiency, Sandeep et.al [13] has implemented a new technique, Indic-Phonetic approach, to obtain the phonetic forms of Hindi and Marathi words.

# III. PROPOSED METHODOLOGY

Complication in the recovery of data is the result of type errors, misspelled words, inconsistent expression habit, and different formats. Matching of words can be defined as the process of determining whether both the words are similar or not. With typographical errors, often there would be interchanging of letters or misspelling of words. For example, ‘Gary’ can be misspelled as ‘Garry’.

Though Soundex and Metaphone are naïve algorithms being used in different applications as embedded tools, each of them have their own disadvantages. Soundex mainly depends on the first letter of the word, has a high overhead in retrieving the near matches, and also it does not consider the phonetic sounds of vowels. Though Metaphone addresses above problems, it only has less accuracy in obtaining the proper matches to the misspelled word. To overcome the limitations in both algorithms, a new algorithm is proposed, namely, Meta Soundex which the high precision as in Metaphone and high accuracy as in Soundex.



**Figure 1: Proposed MetaSoundex Algorithm**

For example, as shown in the Table 1 below, if the word is misspelled in ‘ABACU’ instead of ‘ABACUS’, Metaphone would not be able to obtain the required match. If grouping of letters using Soundex is performed on obtained Metaphone code, then accuracy of the Metaphone algorithm can be improved. Encoding the final code to digits would completely avoid the dependency on first letter.

|  |  |  |  |
| --- | --- | --- | --- |
| **Word** | **Metaphone** | **Soundex on Metaphone** | **MetaSoundex****(Based on proposed alg.)** |
| ABACU | ABK | A12 | 012 |
| ABACUS | ABKS | A12 | 012 |

**Table 1: Phonetic Codes for different Algorithms**

This project also focuses on the development of web application tool which involves choice of algorithm for the misspelled words detected. The experiments in this work will be based on five popular phonetic matching techniques, namely, Soundex, NYSIIS, Metaphone, DMetaphone and Caverphone and the new Meta Soundex algorithm, for the English language. Similarly, the performance of Soundex and Metaphone will be observed for Spanish language along with the new Spanish MetaSoundex. The architectural schema for the proposed web application toolkit is as shown in the Figure 2 below:



**Figure 2: Architectural Diagram of Proposed Web Application Toolkit**

In the proposed methodology, a language detector is used to identify whether the language is either English or Spanish. If the language is Spanish, then the respective Spanish algorithms are used to calculate the phonetic codes. These phonetic codes are compared by retrieving the data from database and approximate matches are obtained. Similarly, for English language, phonetic codes are generated using the above mentioned algorithms and approximate matches are obtained from the database.

The methodology also involves analysis on the performance of the algorithms by calculating the Precision and Recall values. Figure 3 represents the anticipated design of the web application tool.



**Figure 3: Tentative Design of the Web Application Tool**

1. *Evaluation Metrics:*

The performance of phonetic matching algorithms used for information retrieval is evaluated by calculating precision, recall and F - Measure.

*Precision:*

Precision gives the total number of true positives obtained over the total number of suggestions for the obtained true positives.

$P= \frac{\sum\_{}^{}p}{\sum\_{}^{}Number of suggested words for each corrected word}$

, where $p= \left\{\begin{array}{c}1, if the word is corrected\\0, if the word is not corrected\end{array}\right.$

*P* = cumulative precision for an algorithm.

*Recall:*

Recall provides the total number of relevant words over the total number of suggestions [14].

$R=\frac{Number of corrected words }{Total number of misspelled words}$

, where R = recall or efficiency of an algorithm.

The efficiency of an algorithm is obtained by calculating these metrics for different input records.

*F – Measure:*

The F – Measure is calculated based on precision and recall and is defined as the harmonic mean of precision and recall. It is given by

$$F= \frac{2×P×R}{P+R}$$

For the comparison, the maximum F - Measure for different datasets are considered, which vary in size and features.

# IV. PROJECT TIMELINE

The timeline for the project is as shown in figure 4 below. The Research and Background Analysis involves the study of existing algorithms and analyzing their drawbacks to improve the efficiency and performance of the algorithms in both English and Spanish languages. Implementation of the algorithms on real-time data requires collection of all the possible words present in these languages. All the data is collected in the form of dictionary and stored in the database for the analysis and obtaining near matches of the misspelled words.



**Figure 4: Timeline for Project**

# References

1. “Most Widely Spoken Languages in the World”, 2014. [Online]. http://www.infoplease.com/ipa/A0775272.html
2. R. Shah and D. K. Singh, “Analysis and Comparative Study on Phonetic Matching Techniques”, International Journal of Computer Applications Volume 87 – No.9, February, 2014.
3. Beider and S.P. Morse, “Phonetic Matching: A Better Soundex”, March, 2010 [Online].

Available: http://stevemorse.org/phonetics/bmpm2.htm

1. C. Varol and J.R. Talburt, “Pattern and Phonetic Based Street Name Misspelling Correction”, Eighth International Conference on Information Technology: New Generations, 2011.
2. David Hood, “Caversham Project Occasional Technical Paper”, December2004.
3. R. C. Balabantaray, B. Sahoo, S. K. Lenka, D. K. Sahoo, M. Swain, “An Automatic Approximate Matching Technique Based on Phonetic Encoding for Odia Query” IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3, May 2012.
4. A.K. Bhattacharjee, A. Mallick, A. Dey, S. Bandypoadhay, “Enhanced Technique for Data cleaning in text files”, International Journal of Computer Science Issues, Vol. 10, Issue 5, No 2, September 2013.
5. B.P. Pande, H.S. Dhami, “Application of Natural Language Processing Tools in Stemming” International Journal of Computer Applications (0975 – 8887) Volume 27– No.6, August 2011.
6. A.Carstensen, “An Introduction to Double Metaphone and the Principles behind Soundex”, BeyeNETWORK Articles, US Edition, September 22, 2005 [Online]. http://www.b-eye-network.com/view/1596
7. M. P. Angeles, A. E. Gamez, J. G. Moncada, “Comparison of a Modified Spanish Phonetic, Soundex, and Phonex coding functions during data matching process”, 4th International Conference on Informatics, Electronics & Vision (ICIEV), June 2015.
8. F. M. I. Am´on and J. Echeverri, “Algoritmo fon´etico para detecci´on de cadenas de texto duplicadas en el idioma espa´nol,” Ingenier´ıas Universidad de Medell´ın, 2012.
9. A. Mosquera, “Phonetic Indexing with the Spanish Metaphone Algorithm”, February 2012 [Online].

Available: http://www.amsqr.com/2012/02/phonetic-indexing-with-spanish.html

1. S. Chaware and S. Rao, “Analysis of Phonetic Matching Approaches for Indic Languages”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 2, April 2012.

[14] B.A. Kelkar, K.B. Manwade, “Identifying Nearly Duplicate Records in Relational Database.” IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 2, No.3, June 2012.