Computer-based statistical description of phonetical balance for Romanian utterances

A. Cociocanu1, T. Ivânoaica1,2, A.I. Nicolin1,2,* and M.C. Raportaru1

1 Department of Computational Physics and Information Technologies, Horia Hulubei National Institute for Physics and Nuclear Engineering, Reactorului 30, Magurele, Romania
2 University POLITEHNICA of Bucharest, Splaiul Independentei 313, Bucharest, Romania
*Corresponding author: alexandru.nicolin@nipne.ro

Abstract. Motivated by the advent of security solutions which rely on voice biometrics, we revisit by means of extensive computer-based investigations the concept of phonetical balance for Romanian utterances. We show that the standard distribution of phonems offers only a partial description of the phonetics of the language and that more detailed statistical indicators are needed. To this end, we introduce a simple indicator that measures vowel-consonant (or consonant-vowel) sequences and analyze the distribution of consonant clusters for Romanian words. Our results show that the distribution of consonant clusters is scale-free-like (akin to the distribution of words and phrases in large texts) and that large clusters of vowels or consonants are infrequent. This, in turn, indicates that utterances consisting of words which are statistically unrepresentative with respect to the previous indicators are good candidates for benchmarking the efficiency of voice biometrics solutions.

1 Introduction

One of the important changes imposed by the Digital Era concerns the way in which we secure and have access to our assets. Traditionally understood as a physical object that belongs to the owner, the key that grants access to one’s assets has gradually shifted towards something the owner knows and, more recently, towards who the owner is. In fact, the etymology of the word shows the key as a metal piece for opening locks (via Middle English keie cognate with the Middle Low German keie, which means lance or spear), emphasizing the physical nature of the object. The Digital Era, however, has transformed the key into a piece of information, something that (only) the owner knows, usually in the form of an alphanumeric sequence that the owner provides before accessing some digital resources, such as financial data in e-banking systems, medical files on e-health platforms, personal data on cloud storage providers, etc. [1]. In a way, the plethora of digital security solutions, in particular the software ones, such as the hashing methods used to conceal the digital keys, the mechanisms and algorithms used to tunnel and wrap them, the certification systems used to
secure the data transfers, etc., mask the central position that the alphanumeric
digital keys currently have in the information ecosystem. Moreover, although nu-
umerous hybrid authentication methods exist, e.g., the two-factor authentication,
where the addition of a new recognition method is improving the authenticity of
the owner’s identity [2], we observe that one component of this scheme remains
invariant: the (tried and tested) alphanumeric key.

Naturally, the assets themselves have shifted towards the digital realm as well,
as have many of our activities. In fact, digital technologies have substantially
changed the way we socialize and entertain ourselves, the way we work and
learn, and so on, and we now have a new discipline, namely digital sociology,
solely dedicated to the way these technologies are impacting our everyday life
[3].

Finally, the past decades have placed more emphasis on (implicitly auto-
mated) biometrics security solutions, which allow the owner of some assets to
access them not by using something that he possesses or something that he
knows, as done in the past, but rather by identifying himself as the owner [4].
The term biometrics itself is defined as “automated recognition of individuals
based on their behavioral and biological characteristics” (see ISO/IEC JTC1
SC37), the most common biometric traits used for authentication of users being
the fingerprint, the face, the iris, the palm-print, the retina, and the voice. In
fact, any human physiological and/or behavioral characteristic can be used as
a biometric characteristic, as long as it satisfies a series of requirements such
universal, i.e., every person using the system should possess the trait, unique-
ness, i.e., the trait should be sufficiently different for individuals in the relevant
population such that they can be distinguished from one another, etc., but at
the moment the voice is one of the most used traits [4].

As part of building a voice-biometrics identification system which is largely
text-independent (i.e., no pass phrase) and shows little sensitivity to ambiental
noise [5], we revisit by means of extensive computer-based investigations
the concept of phonetical balance for Romanian utterances. The goal of our in-
vestigation is to have statistical descriptors of the phonology of the Romanian
language that will be helpful in the development stages of the aforementioned
voice-biometrics identification system. To this end, we go beyond the standard
distribution of phonems and analyze the distribution of consonant clusters for
Romanian words to identify the most important ones. Moreover, we propose a
simple indicator that measures vowel-consonant sequences to show that large
clusters of vowels or consonants are infrequent.

\section{Distribution of phonems}

While the mathematics behind the distribution of phonems in a given text is
relatively simple, the main technical challenge comes from finding a set of texts,
usually very large, that are representative for the language under scrutiny. In the
case of Romanian language [6] the text was acquired using the Web-as-resource or
Web-as-corpus approach (considering as sources mainly online Romanian news-
papers and transcripts of the discussions in the European parliament) [7], which produced more than 9 million phrases, the largest Romanian plain text corpus to date [7]. The results of this analysis clearly show that the 34 phonems identified are qualitatively different, some of them being very common, while others somewhat infrequent. In fact, the first six phonems correspond to more than 50% of the entire phonem usage, while the last six phonems have an occurrence frequency in between 0.27% and 0.03%.

The aforementioned distribution of phonems (complemented with related results in Ref. [8]) gives an accurate global description of the Romanian language which can be now compared with languages (such as English and French) for which such statistical descriptions have a long history. We note, however, that in the case of smaller texts this statistical description is insufficient, as two small texts of similar if not identical distribution of phonems may be, in fact, substantially different from a phonetic point of view. As an elementary example we note here two simple Romanian sentences, namely S1: “Oaia e proastă” (in English: The sheep is dumb) and S2: “Ia ta e poroasă” (in English: Your blouse is porous), which consist of the same phonems but differ considerably. The first sentence, for instance, has a series of four subsequent vowels (namely “oaia e”) and a consonant cluster of two letters (namely “st”), while the second has no consonant cluster and the longest series of vowels is of length two (namely “ae” and “oa”).

In the language of statistical physics, if the statistical ensemble is very large, then the distribution of phonems becomes the key descriptor of the text under scrutiny, as all other features (say, vowel and consonant clusters, distribution of words, and so on) average out. For short texts, however, the distribution of phonems should be complemented with additional information about the vowel-consonant sequences and the types of consonant clusters.

### 3 Distribution of consonant clusters

Motivated by the previous example we embarked on a detailed statistical study of the Romanian vocabulary using the database of Dexonline [9], which is an open source collection of the main dictionaries of the Romanian language. For our analysis we used a set of more than 90,000 words, which roughly correspond to the lexicon in Ref. [10], which is the main dictionary of the Romanian language.

Through a thorough statistical analysis we identified all two-consonant clusters, independent of their position in a word, and ranked them according their occurrence race. To simplify the analysis, the results for the letter s include also the letter s, while those for the letter t also include the letter t. The bubble-plot in Fig. 1 shows the main consonant clusters, indicating their occurrence through the size of the bubble. The main message of the plot is that there are a few frequent consonant clusters (such as “st”, “nt”, “tr”, “pr”, etc.) which appear in 5% to 10% of all words in the Romanian vocabulary and numerous other infrequent clusters (such as “lb”, “sf”, etc.). A rapid inspection of the plot shows the tendency to have consonant clusters using the letters in the second half of
Fig. 1. Bubble plot of the main two-letters consonant clusters observed in Romanian words. The size of the bubble is proportional to the number of words in which the cluster appears.

the consonant series. Moreover, there is a clear assymetry with respect to the first diagonal, meaning that the frequency rates of a given cluster and its inverse are substantially different. To understand this property let us look at the “tr” cluster which is more frequent than the “rt” cluster, or, more clearly at the “st” and “ts” ones. The first order of the consonants, “st”, corresponds to one of the most important Romanian cluster, while “ts” is inexistent.

What is more interesting is that the distribution of the consonant clusters follows a free-scale-like distribution. Taking $P(k)$ as the probability that a given cluster appears in $k$ words, we observed (see Fig. 2, the upper panel) that $P(k) \approx k^{-\gamma}$ where $\gamma \approx 3.2$. This statistical behavior shows a striking resemblance to the so-called Zipf law which states that the frequency of a given word is inversely proportional to its rank in the frequency table [11]. A classical textual econometrics study on the Brown Corpus of American English showed “the” as the most frequent word in the vocabulary (with an occurrence rate of almost 7%), “of” as the second most frequent word with an occurrence rate of roughly 3.5%, etc. Similarly, it has been shown that the distribution of word sequences (the so-called $n$-grams) follows the same pattern, provided the reference texts are large enough [12]. The key feature of the Zipf law is that on a log-log plot the distribution is linear with a negative slope, which is similar to what we noticed in Fig. 2 (the upper panel). In our case “st” was the most frequent consonant cluster with a frequency rate of 9.4 %, “nt” was the second most frequent cluster (8.5 % frequency rate), “tr” was the third one (7 % frequency rate), etc. This distribution is typical to many systems, ranging from social networks (such as the collaboration of movie actors in films and the co-authorship of papers),
the internet, the protein-protein interactio, etc. [13]. While Zipf’s law has been verified in numerous contexts, the mechanisms behind it remain largely elusive, despite numerous models which capture some of its features. Zipf, for example, understood the law through the principle of least effort, which has been often revisited by means of advanced mathematical models [14], while others consider the preferential attachment mechanism which basically says that the speakers tend to use some words more often than others [15].

It is tempting to see the correlation between the frequency rates of the consonant clusters and their etymology, but such an analysis is of little insight, as the Romanian dictionaries only record the language from which a given word entered into the Romanian vocabulary, and not the language of origin [16]. Finally, we note that the distribution of three- and four-letters consonant clusters brings only a small correction to the aforementioned statistics and that a detailed study will be reported elsewhere.

![Graph](image)

**Fig. 2.** (Upper panel) Distribution of the probability that a given cluster appears in a given number of Romanian words. Please note that the plot is log-log. The dashed line shows the linear fit (in log-log axes) of the function and corresponds to a slope of $-3.2$. (Lower panel) The same distribution as in the upper panel on a graph without log-log axes. Please notice the rapid decay of the distribution for a number of words larger than 2000.
4  Phonetical balance

The previous discussion on consonant clusters brings some clarification to the phonology of Romanian language, but an instrument is needed to quantify the vowel-consonant sequences. To this end, we introduce for each word in the vocabulary the function

\[ m = \frac{1}{n-1} \sum_{j=1}^{n-1} l(j)l(j+1), \]

where \( n \) is the number of letters of the word and \( l \) is a boolean function equals to 1 if the \( j \)-th letter is a vowel and \(-1\) if the \( j \)-th letter is a consonant. Please note that a normalization factor equal to \( 1/(n-1) \) has been introduced in the definition of \( m \) such that from the values of \( m \) one can directly compare words of different lengths. For words consisting of perfect sequences of vowels and consonants such as “calamitate” (in English: calamity), “repetare” (in English: repetition), “sare” (in English: salt), etc., the \( m \)-function is numerically equal to \(-1\), while for the (admittedly fewer) words consisting almost entirely of vowels or consonants, such as “oaie” (in English: sheep) and “ouă” (in English: eggs) on the vowel side and “strâmb” (in English: crooked) and “prompt” (in English: prompt) on the consonant side, the \( m \)-function is always larger than 0. In fact for “oaie” and “ouă” the \( m \)-function is equal to exactly +1 but this is valid for a very short list of words.

In Figs. 3 and 4 we show the distribution of the number of Romanian words as function of the parameter \( m \), using a histogram plot with different number of bins, to show that the majority of words corresponds to negative values of \( m \) (\( \bar{m} \), the average value of \( m \) being, in fact, very close to \(-0.5\)), thereby indicating that most Romanian words are structured as (slightly) imperfect vowel-consonant sequences. Fig. 4, in particular, shows extremely clearly that there is a large set of words, around 18.5 % of the investigated vocabulary, which corresponds to \( m = -1 \).

This shows, incidentally, that the first sentence discussed in Sec. 2 consists of words from the tail of the distribution of \( m \) (though the consonant cluster is one of the very frequent ones), while the second sentence consists of words from the bulk part of the distribution. Let us also note that approximately 26 % of the Romanian words correspond to \( m < -0.75 \). Finally, let us mention that it is very tempting to compute the global \( m \) value considering the reported frequencies of Romanian words [17], but the computations were done before the advent of the computer and the validity of some of the reported frequencies has been questioned [16]. We can, however, use equation (1) for entire sentences, just like for words, and obtain \( m_{S1} = 3/11 \) for the first sentence and \( m_{S2} = -5/11 \), for the second one, which indicates that the first sentence is less representative in a statistical sense with respect to the vowel-consonant (or consonant-vowel) sequences than the second one.
Fig. 3. Distribution of the number of Romanian words as function of the parameter $m$ introduced in equation (1). Please note that because of the low number of bins in the histogram it appears, quite erroneously (see Fig. 4), to have a continuous transition from one set of $m$-values to another. The plot is, however, extremely useful because it allows an easy assessment of the average value of $m$, which is around $-0.5$.

Fig. 4. Detailed distribution of the number of Romanian words as function of the parameter $m$ introduced in equation (1). Please note the isolated cluster close to $m = -1$ and the bunch in between $m = -0.8$ and $m = 0$. From the graph one can easily assess that there are around 17,500 words with an $m$-value very close to $m = -1$. 
5 Conclusions

In this paper we have analyzed by computational means the phonetical balance of Romanian words and introduced two indicators that go beyond the standard distribution of phonems. We have shown that the distribution of consonant clusters in Romanian words obeys a scale-free-like distribution and that large clusters of vowels or consonants are infrequent. The distribution of consonant clusters is similar to the well-known Zipf law that gives the distribution of words and short sentences in that it shows that there are a few very frequent consonant clusters and numerous others which are considerably less frequent. Our results suggest that a reliable voice-based biometrics solution should be particularly benchmarked against utterances which consist of words with infrequent consonant clusters and words with positive $m$-values, as their statistical unrepresentativeness makes them good candidates for identifying the flaws of a given biometrics solution.

As a natural extension of this work we intend to refine the current results by taking into account the position of a consonant cluster with respect to the syllables of a word. Moreover, future research should be focused on a global indicator (such as the Shannon entropy) which considers not only the relations between the nearest neighbours letters, but also long-range in-word correlations between letters and clusters.

Acknowledgement The work was supported by the Horizon 2020 SpeechXRays project. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 653586. The authors thank Dan Caragea and Mihai Bărbulescu for insightful discussions.

References

1. Bonneau, J., Herley, C., van Oorschot, P.C., Stajano, F.: Passwords and the Evolution of Imperfect Authentication, Comm. ACM 58, 78-87 (2015).
2. See, for example, Apple’s Touch ID, the biometric fingerprint plus PIN code, the Google 2-Step Verification, and the user telephone number plus user password.
3. Lupton, D.: Digital Sociology, Routledge (2014).
4. Jain, A.K., Flynn, P., Ross, A.A., Eds.: Handbook of Biometrics, Springer (2008).
5. See, in particular, the SpeechXRays project (http://www.speechxrays.com) which “will develop and test in real-life environments a user recognition platform based on voice acoustics analysis and audio-visual identity verification” which uses “text-independent speaker identification (no pass phrase)” and “low sensitivity to surrounding noise” (as retrieved on June 2016). The testing will be done in Greek and Romanian.
6. See a detailed description of the Romanian language on its Wikipedia entry at https://en.wikipedia.org/wiki/Romanian_language.
7. Stanescu (Pasca), M., Cucu, H., Buzo, A., Burileanu, C.: ASR for Low-Resourced Languages: Building a Phonetically Balanced Romanian Speech Corpus, 20th European Signal Processing Conference (EUSIPCO 2012), EURASIP 2012, ISSN 2076-1465.
8. Stanescu (Pasca), M., Buzo, A., Cucu, H., Burileanu, C.: Statistical Phonetic Analysis of the Romanian Language for Speech Recognition and Synthesis Tasks, 54th International Symposium ELMAR 2012.

9. The Dexonline dictionary is publicly available at https://dexonline.ro and consists of more than fifty distinct dictionaries (see the complete list at https://dexonline.ro/surse) which have a very broad coverage. The dictionaries go from general-use prescriptive dictionaries and thesauruses to topical and orthographic dictionaries.

10. Explanatory Dictionary of the Romanian Language (in Romanian), Romanian Academy, “Iorgu Iordan” Linguistic Institute, Editura Univers Enciclopedic, second edition, 1996.

11. Zipf, G.K.: Human Behavior and The Principle of Least Effort. An Introduction to Human Ecology, Addison-Wesley Press (1949).

12. Ha, L.Q., Sicilia-Garcia, E.I., Ming, J., Smith, F.J.: Extension of Zipf’s Law to Words and Phrases, Proceedings of the 19th International Conference on Computational Linguistics - COLING (2012).

13. Barabasi, A.-L., Reka, A.: Emergence of scaling in random networks, Science 286, 509-512 (1999).

14. Cancho, R.F., Sole, R.V.: Least effort and the origins of scaling in human language, Proc. Nat. Acad. Sci. 100, 1788-791 (2003).

15. Lin, R., Ma, Q.D.Y., Bian, C.: Scaling laws in human speech, decreasing emergence of new words and a generalized model, arXiv:1412.4846v2 (2015).

16. Dinu, M.: Personalitatea Limbii Române (in Romanian), Cartea Românească (1996).

17. Juillard, A., Edwards, P., Juillard, I.: Frequency Dictionary of Romanian Words, Mouton & Comp. (1965).