

Content Discovery Using Perceptual Automation

Invited Paper

Alexander Iliev

University of Wisconsin
Stevens Point, WI 54481, USA
IMI - BAS
8, Akad. G. Bonchev, Str.
1113, Sofia, Bulgaria
al.iliev@math.bas.bg

ABSTRACT

In this work, an innovative media content discovery and selection system is proposed based on human behavior. There were two parts to this project: first, in the perception phase, a speech signal is used for analysis with the intention to detect and extract emotional cues; and second, in the automation phase, textual information was used to determine the sentiment using natural language processing methodology. The speech cues are first captured from the signal and then classified in three emotional categories: upbeat, positive or happy; emotionless, flat or neutral; and downbeat, negative or sad. Then in order to match the collected emotions from speech to specific contextual information captured from a real source, using natural language processing techniques five books from The Game of Thrones were processed. The results were summarized and mapping between the two systems was drawn so that books can be chosen based on the degree of each emotion portrayed from speakers. This can be done in a non-intrusive fission without direct and explicit human-machine intervention.

CCS CONCEPTS

Human computer interaction (HCI) - HCI theory, concepts and models; Interactive systems and tools, User interface management systems; Collaborative and social computing, Social recommendation.

KEYWORDS

Content Discovery Selection Emotion Recognition Speech Glottal Processing NLP Perceptual Automation

ACM Reference format:

Alexander Iliev. 2018. Content Discovery Using Perceptual Automation. *In Proceedings of the 10th International Conference on Management of Digital EcoSystems (MEDES '18)*, September 25–28, 2018, Tokyo, Japan. ACM, New York, NY, USA, 6 pages.
<https://doi.org/10.1145/3281375.3281399>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
MEDES'18, September 25–28, 2018, Tokyo, Japan
© 2018 Copyright held by the owner/author(s). 978-1-4503-5622-0/18/09...\$15.00

INTRODUCTION

Emotions vary in the way they are depicted. We know from psychology that there are many emotional states that can be observed, but detecting them and applying them in real-world applications is another matter that very much depends on the practical implementation and the problem to solve. In data mining, before even collecting, cleaning, and normalizing the data we want to work with, we always have to know the type of data we will be dealing with so that we can determine the best approach and specify number of methodologies to prepare it for analysis. In the case of this work, there are two types of data of interest to us: speech and text. The idea was simple: to create a working system that can discover and propose possible books to read or audio books to listen to based on human behavior. This could be further extended to recommend any type of media: video, images or audio, that can be used as entertainment based on media's metadata. If we consider a simple use case where one is seeking a source of entertainment as described here, the knowledge gathered from the source can tremendously simplify the task of dealing with thousand of choices when it comes to selecting an appropriate media that corresponds to one's mood. If we are tired from the working day for example we might want to choose something calm and soothing that does not require the use of too much effort. In this case, a movie that has a slower pace and is more downbeat can be recommended for the audience. In another use case where the audience is full of energy for example, a cheerful and more upbeat music or video clips can be chosen automatically based on the perceptual cues extracted from the speakers speech. That is so, because we often seek entertainment based on our current mood and environment, which makes these use cases valid in an every-day life scenario.

When it comes to selecting the source of behavior for analysis, we only have few choices: using video, images or audio recording from speech. The first two types of media are great sources of information because many cues can be extracted from image processing alone. That is because of the way we portray emotions. Emotions can be collected in various different ways. Examples are: facial expression, body language, physiological cues such as sweat, skin coloration, etc. One of the problems with cameras however is that by nature they are more intrusive and may easily obstruct privacy. Another is purely technical since a system that consists of array of cameras is harder to implement and is more

costly. Processing video information from such systems can add to the computational strain and is harder to implement. It is also a matter of positioning in terms of the subject or the camera. In order to capture someone's facial expression, that person's face has to be in the direction of the camera and they cannot be turning their back to it. That alone may require the installation of multiple cameras. All of these make the collection of emotional cues from video and images very challenging. In these terms, speech alone looks more attractive since collecting audio signals does not have these issues. The physical nature of sound helps us so that it propagates in all directions equally and makes the need for someone to be in front of the data collection point (microphone) nonexistent. Moreover, it is easier to use and setup a single microphone, which alleviates the computational strain and investment.

When processing speech we need to also focus on the techniques used for feature extraction. Then we need to consider the language in which the emotion is portrayed since emotions are culturally dependent, in some cases, accents and gender can also be of significance when it comes to precision. All in all the system described here generalizes all of these parameters so that it focuses on the overall emotional content as gathered from a specific environment with multiple users over some time. This alone makes the system user, gender and context independent, hence applicable in everyday life settings and completely defines the first portion of this work – the perception phase.

Considering the second portion of this work, that was the automation, the media of choice was books in an electronic text format. After some processing, they were matched to the emotional content extracted from phase one. Natural language processing methodologies were applied to the texts in order to determine the overall emotional content from each book. Once this process was complete, it was a matter of technicality of how the user of the system would have liked to map the first and second part of this system and fine-tune its hyper parameters in order to make it work. For this portion, five books in text format from The Game of Thrones were used for processing with the intention to extract relevant mood states from each book as a whole. The moods corresponding to the three emotional states chosen for analysis with degrees from [<0 , 0, >0] or [positive, neutral, negative] were [happy, neutral, sad].

SPEECH SIGNAL PROCESSING

The emotional content in speech signals is solely contained in the voice parts of the speech signal and in any part that involves the use of the glottis. To put it in simple terms, the glottis is the part that generates a pulse-like train signal that propagates through the vocal tract and gets an additional sound coloration specific for each individual. That is so because of the resonances generated from the specific acoustic environment of the vocal tract and is different for each gender as well as each person. The system can simply be decomposed in three main parts as follows:

$$A(z) = L(z)V(z)G(z), \quad (1)$$

such that, $A(z)$ is the speech produced from the speaker, $L(z)$ is the radiation at the lips, $V(z)$ is the transfer function of the vocal tract and $G(z)$ is the model of the glottal signal. To obtain the glottal waveform from speech, firstly the lip radiation function needs to be calculated. This is done such that for $L(z)$, a simple 1st order filter is constructed or:

$$L(z) = 1 - \alpha z^{-1}, \quad (2)$$

where $0.95 \leq \alpha \leq 1.0$. The vocal tract needs to be obtained next. If the radiation of the lips and the vocal tract function can be sufficiently exhibited from the composite speech signal, it follows that the inverse filtering of the glottis can be performed reliably for estimating the model of the glottal pulse. When voicing occurs the vocal tract can be estimated as an all-pole filter that takes the form shown in Eq. (3) below:

$$V(z) = \frac{1}{1 + \sum_{i=1}^p a_i z^{-i}}, \quad (3)$$

where, p is the prediction order. The filter coefficients for $V(z)$ can easily be acquired through applying a linear prediction (LP) analysis methodology by using the autocorrelation or covariance methods as explained by Rabiner and Schafer [1]. The next and final step is to solve the expression for the glottal signal $G(z)$ as shown in Eq. (4) below, which we refer to as - inverse filtering approximation of the glottal signal, or:

$$G(z) = \frac{A(z)}{V(z)L(z)}, \quad (4)$$

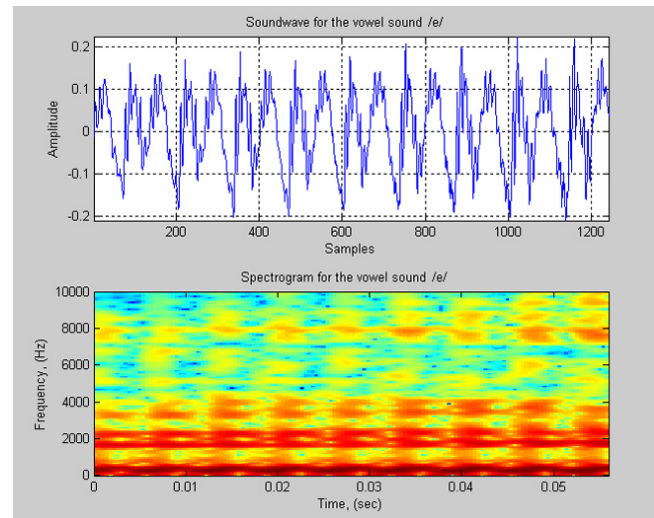


Figure 1: Spoken vowel sound 'e': waveform (top) and spectrogram (bottom)

Figure 1 shows the sound wave for the vowel sound 'e' on the top, along with its corresponding spectrogram at the bottom. As expected the most energy and information of interest to us is

located at the lower frequency spectrum. Inverse filtering is applied to all vocal regions in the speech signal and is described next.

INVERSE FILTERING

There are number of method proposed for inverse filtering of the glottal signal from speech and they vary in the way the speech was recorded. One method suggests that the signal has to be recorded in the mouth [2], thus eliminating the need to account for $L(z)$, which would simplify the calculations. However, this is not very practical since it is a very intrusive procedure and is not realistic for a real-world solution. Therefore, methods that deal with a regular recording away from the mouth are considered here. One such method is the Sequential Covariance Method as proposed by Wong et. al. [3]. This method is using an all-pole model for $V(z)$ when a vowel is produced. It is employing a covariance analysis for the estimation of the vocal tract's transfer function. This is recorded and estimated in the glottal closure phase during speech production. That was determined using the resulting normalized linear prediction error energy waveform and a typical result is shown in Figure 2 below:

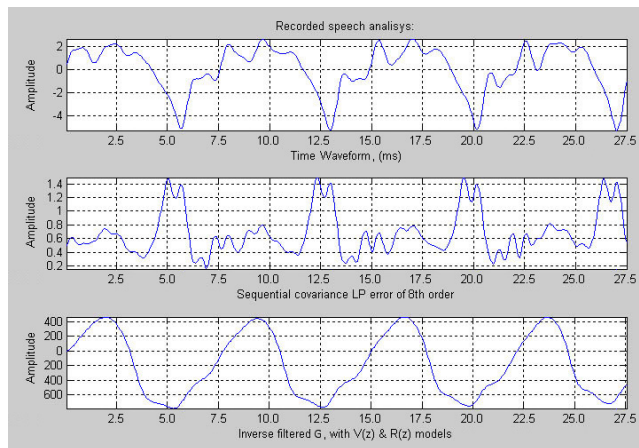


Figure 2: Recorded speech signal and the inverse filtered glottal waveform (bottom) [4]

The procedure of inverse filtering removes the radiation of the lips, cancelled by integrating the output of the inverse filter. One observation is that each phonation type has a different acoustic coloration and as a result the inverse filtering process produces a unique glottal flow typical for that phonation. Furthermore, this method shows that this covariance analysis provides a least-squares estimation of the vocal tract all-pole model $V(z)$ as seen in eq. (3). To obtain the analysis filter we use:

$$F(z) = \sum_{i=0}^M a_i z^{-i} \quad (5)$$

where, $a_0 = 1$ and M is a factor that represents double the number of formants in the speech signal $s(n)$ [4]. The error residue signal is obtained after $s(n)$ passes through $F(z)$ and it looks like this:

$$\varepsilon(n) = s(n) + \sum_{i=0}^M a_i s(n-i) \quad (6)$$

It follows that the filter coefficients for the filter $F(z)$ are obtained by minimizing the squared error $\alpha_M(n)$, represented by:

$$\alpha_M(n) = \sum_{j=n}^{n+N-M-1} \varepsilon^2(j) \quad (7)$$

In eq. (7), the window length is represented by N number of samples. The squared error $\alpha_M(n)$ is calculated by shifting the window on a sample-by-sample basis. $F(z)$ is thus obtained via a linear prediction covariance method, applied over a window width that is convolved with the input speech signal s , starting at $(n-N)$ and ending at point $(n+N-M-1)$. Closures are found when $\alpha_M(n) = 0$. Openings are determined when the first following sample different than zero is detected at n_2 or at point $(n_1+N-M-1)$. To avoid bias by external system gain, system error is normalized as follows:

$$\eta(n) = \frac{\alpha_M(n)}{\alpha_0(n)} \quad (8)$$

where, $\alpha_0(n)$ represents the input signal energy. One of the best advantages of this procedure is that its simplicity requires less computational power. It also detects glottal closures more accurately than it detects openings. Hence, in order for the openings to be represented more accurately a pre-emphasis filter of the signal has to be used before calculating the LP covariance. Furthermore, to make this method speaker independent as well as robust to noise, another techniques has been used for the glottal signal, namely the use of Glottal Symmetry (GS). It represents the ratio between the opening and closing phases of the glottis as shown in Figure 3:

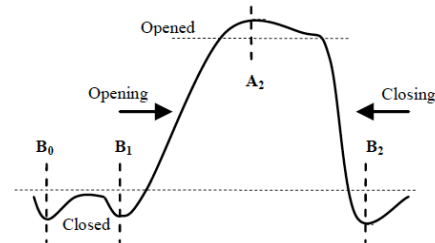


Figure 3: Glottal signal showing all phases: opening, opened, closing and closed phases [4]

Thus, GS is determined as follows:

$$GS = \frac{\text{Closing}}{\text{Opening}} = \frac{|A2-B2|}{|B1-A2|} \quad (9)$$

DETECTING EMOTIONS

One of the most important qualities of using signals recorded by microphones alone is the privacy and non-intrusiveness when the signal is collected. For that reason speech sounds can be captured in a controlled environment and then can be labeled for different emotions of choice. As already mentioned, for our task we chose some of the three most basic emotions [happy (H), neutral (N), sad (S)]. This is a common preset when collecting the signals for training and testing. It is vital that the speech is collected without intrusion in order for the speakers in the room to unintentionally and indirectly talk to the microphone.

Different collections of behavioral characteristics can be collected in this fission, which can later be gathered in vectors representing different features [fn – fnm]. After each feature is collected from multiple observations, they can be repackaged and emotional matrices can be organized, such that the later can be represented as a behavioral matrix for each emotion:

$$\mathbf{E}_{Sp}^{HNS} = \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_2 & \dots & \mathbf{f}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{f}_{1m} & \mathbf{f}_{2m} & \dots & \mathbf{f}_{nm} \end{bmatrix}, \quad (10)$$

where, each column in matrix ESp from eq. (10), depicts a different attribute in feature space directly extracted from the speech signal and symbolized by f. After this step was complete, the data had to be cleaned and normalized before it was further processed for training. Equation (10) depicts the general case when more than one feature is collected, however for our case we only deal with the collection of the glottal signal through inverse filtering and subsequently we use the GS for further analysis, training and testing. Therefore, for our case the matrix takes the form:

$$\mathbf{E}_{GS}^{HNS} = \begin{bmatrix} \mathbf{f}_{H_1} & \mathbf{f}_{N_1} & \mathbf{f}_{S_1} \\ \vdots & \vdots & \vdots \\ \mathbf{f}_{H_m} & \mathbf{f}_{N_m} & \mathbf{f}_{S_m} \end{bmatrix}, \quad (11)$$

We can now calculate the probability of certain behavior to occur in f by using the probability density function (PDF):

$$\int_a^b \mathbf{f}(\mathbf{x}) \mathbf{dx}, \quad (12)$$

where, a and b are boundary conditions found for each of the three emotional states such that $\mathbf{x} = \text{HNS}(1 \div m)$ takes a value from the set E of extracted features as shown in eq. (11) [5]. The following conditions must be met in order to have a valid PDF:

$$\mathbf{f}(\mathbf{x}) \geq 0 \text{ for } \forall \mathbf{x}, \text{ and } \int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x}) \mathbf{dx} = 1, \quad (13)$$

With this expression we set the condition in which the probability density function can determine the probability that f lays between a and b for each of the three emotions, or:

$$\mathbf{P}(\mathbf{a} \leq \mathbf{f} \leq \mathbf{b}), \quad (14)$$

After the entire process has been described and set, all we had to do was to collect a variety of speech samples in a text, speaker and environment independent setup so that we could further train the system with a healthy variation of samples representing the three emotional states of interest. We then organized all of these samples in a matrix containing the feature vectors for the specific human behaviors sought, and after cleaning and normalizing these samples were sent to the classifier for training. This, as expected is a supervised machine-learning problem and is explained in more details in [6], [7].

NATURAL LANGUAGE PROCESSING

In the second phase of this work, that is the automation, the results from the speech sample training process had to map into the same emotional set we chose [H, N, S] as extracted from text in our case, or metadata in every other practical implementation. This was necessary in order to connect the results from the perceptual phase to the automation phase and thus making the system complete. For that reason, five books in a UTF-8 ISO-8859-1 format from The Game of Thrones in plain text were used for processing by means of a sentiment analyzer from a Natural Language Processing Toolkit (NLTK) for Python. More specifically the Vader module was used as described in [8]. The number of sentences and words statistics aggregated for all five books are displayed in Table 1 below:

Books stats	
Number of sentences	145,015
Number of words	2,146,756

Table 1: Accumulated stats collected for all five books

For each book, each sentence and every word were tokenized so that a polarity count determined the emotional content of words in a subsequent step. Table 2 depicts the total aggregated polarity based on the three emotional classes H (positive), N (neutral) and S (negative) and Figure 4 represents these results graphically.

Total Polarity Count			
	H - positive	N - neutral	S - negative
Book 1	6296	14494	6452
Book 2	6930	16017	6945
Book 3	9029	19182	9393
Book 4	6074	8840	5829
Book 5	8286	12854	8397

Table 2: Total sentiment count for each of the five books

Since these results depict the overall sentiment feeling from each book, it was found pertinent to make the polarity search based on specific words or characters used in the story plot of the books. This is why, a following analysis was performed based on the top 10 characters in the story, and based on that three were randomly chosen to be used for the next more specific step. These characters are shown in Figure 5.

At this point, the sentiment recognition was narrowed further down based on specific keywords in the metadata. In our case, the three character names were chosen randomly and were: "Jon", "Daenerys", and "Bran".

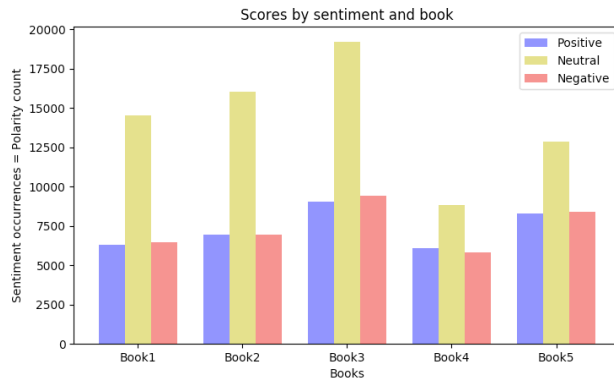


Figure 4: Total sentiment count for each of the five books

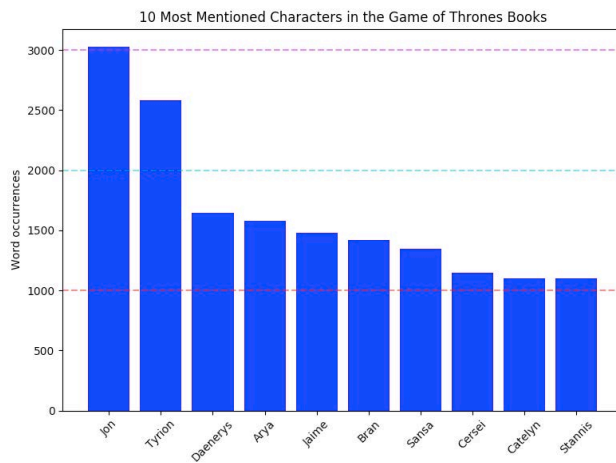


Figure 5: Top 10 most mentioned characters in all five books

In the next step, all the sentences that contained these words were processed in all of the five books. Each of the sentiments captured from the books, given the three keywords, are depicted in Table 3, and have their corresponding values displayed in Figure 6:

	Polarity Count		
	H - positive	N - neutral	S - negative
Book 1	18	36	17
Book 2	18	30	25
Book 3	33	47	42
Book 4	12	20	18
Book 5	10	27	33

Table 3: Sentiments for three selected characters extracted from each of the five books

The lexical dispersion plots for the three characters of choice are displayed in Figure 7 and are inline with the findings from Figure 5. This plot basically shows the locations of these words in the

context of the specific book from the beginning to the end. In this way we could further investigate changes in the text over time. This simple count or analysis can be further applied for each of the five books.

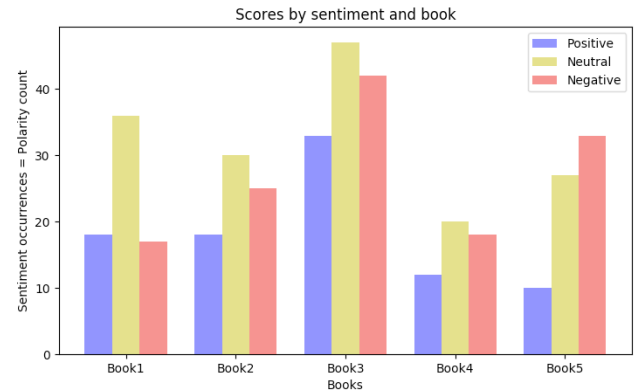


Figure 6: Sentiments for three selected characters extracted from each of the five books

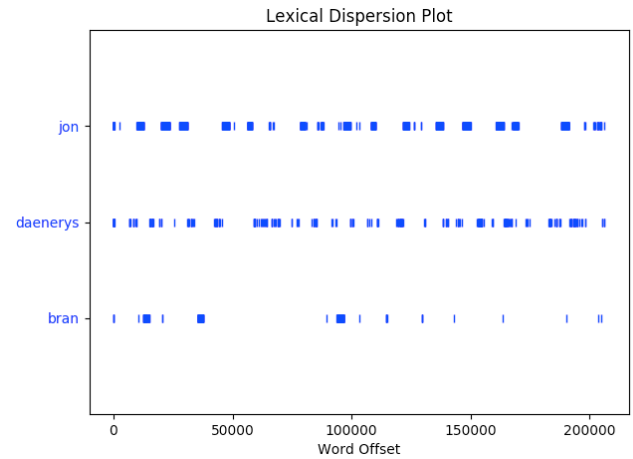


Figure 7: Lexical dispersion plot from book five for three selected characters

It is important to note that the sentiment classification used here, counted the occurrences of the positive, neutral and negative words in any given sentence and then made a classification based on polarity weight. The Vader module returned a float for sentiment strength in each sentence.

CONCLUSIONS

In this work a practical approach for system control and automation was proposed based on perceptual cues in the speech emotion domain. One of the most important parts was to develop the mapping interface between the two stages subjects to this discussion, namely the perception and automation stage. The mapping between the two was achieved by finding a common ground in the emotion space by using three emotional states H (positive), N (neutral) and S (negative). The idea was that machines could learn the behavioral patterns of different speakers and then decide what type of content can be recommended

automatically. From the second portion of the study we see that sentiment can be drawn as a general feeling for the entire media books in this case, which was shown in Figure 4. There we notice that Neutral prevails in each book, which is expected and is consistent with other research [4]. In this case if the emotion extracted from speech was more positive or Happy, then book 4 can be automatically recommended since polarity of the same type prevails there. Furthermore, we can narrow down the system search based on personal preferences or personalization, such that we can make the same conclusion based on specific pre-selected keywords, or characters in our case. Such is the case displayed in Table 3 and Figure 6. It can be seen that for the same emotion as in the previous example, or Happy, we can now recommend book 1 from that same set. The occurrences in time can be further validated by the use of Lexical Dispersion Plot as shown in Figure 7. By extracting emotional content from speech in a robust way while using the techniques discussed in sections 2 (*Speech Signal Processing*), 3 (*Inverse Filtering*), and 4 (*Detecting Emotions*), we can map them to any system that uses the same emotional content. The NLP approach used in this work was a good example of how this can be achieved.

FUTURE RESEARCH

To take it a step further this research can be implemented in even more sophisticated setup, one that extracts cues from the media itself without the need to have a proper metadata attached to it. For that purpose a multimedia approach can be taken such that attributes can be extracted from videos, images, speech and

text independently, or in any combination of the above that is related to the targeted media.

Other features extracted from the perceptual phase can also be added to the feature vectors, but this should be determined based on application, needs, computational throughput, system cost, etc.

REFERENCES

- [1] Rabiner L. and Schafer R., 1978. Digital Processing of Speech Signals, Prentice Hall.
- [2] Rothenberg M., 1973. A New Inverse-Filtering Technique for Deriving the Glottal Air Flow Waveform During Voicing. Journal of the Acoustical Society of America, Vol. 53, pp. 1632-1645.
- [3] Wong D., Markel J., and Gray A., 1979. Least Squares Glottal Inverse Filtering from the Acoustical Speech Waveform. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No. 4, pp. 350-355.
- [4] Iliev, A.I., "Emotion Recognition From Speech", Monograph, Lambert Academic Publishing, 2012.
- [5] Stanchev, P.L., Marinova, D.P., Iliev, A.I., "Enhanced User Experience and Behavioral Patterns for Digital Cultural Ecosystems", The 9th International Conference on Management of Digital EcoSystems (MEDES'17), Bangkok, Thailand, 7-10.Nov.2017, ACM, ISBN:978-1-4503-4895-9, pp. 288-293
- [6] Iliev, A.I., Scordilis, M.S., "Spoken Emotion Recognition Using Glottal Symmetry", EURASIP Journal on Advances in Signal Processing, Volume 2011, Article ID 624575
- [7] Iliev, A., and Stanchev, P. 2017. Smart multifunctional digital content ecosystem using emotion analysis of voice. In Proceedings of the International Conference on Computer Systems and Technologies CompSysTech'17 (Ruse, Bulgaria).
- [8] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.