Automatic recognition of emotions in the description of motion pictures for television archives

Jorge Caldera-Serrano
https://orcid.org/0000-0002-4049-1057
jcalser@unex.es
Universidad de Extremadura

Resumen

Los departamentos de información de las redes de televisión están experimentando una automatización continua de sus procesos documentales. Bajo el marco de la descripción de la imagen, existe la posibilidad de implementar técnicas e imágenes de Reconocimiento Automático del Habla (ASR), que pueden usarse para identificar las diversas emociones que se pueden ver en las imágenes. Los periodistas requieren información de información connotada y que también está presente explícitamente a través de gestos faciales y vibración de voz, por lo que los departamentos de información audiovisual deben describir esos elementos o encontrar herramientas automatizadas para ayudar a identificarlos. En este artículo, presentamos el método y la validez de automatizar los procesos para extraer información de las emociones utilizando técnicas biométricas. Para esto, realizamos una revisión bibliográfica y visitamos centros de información de televisión para determinar los requisitos y luego capturar los cambios necesarios en los mecanismos de automatización del sistema.

Palabras clave

Reconocimiento de emociones / automatización de documentos / documentación audiovisual digital / Archivos de televisión / Biometría / Reconocimiento de video / Reconocimiento de audio

Abstract

The information departments of television networks are undergoing a continuous automation of their document processes. Under the framework of image description, there is the possibility of implementing Automatic Speech Recognition

(ASR) techniques and images, which can be used to identify the various emotions that can be seen in images. Journalists require information from connoted information and that is also explicitly present through facial gestures and voice vibration, so audiovisual information departments must describe those elements or find automated tools to help the identification thereof. In this paper, we present the method and validity of automating the processes for extracting information from emotions using biometric techniques. For this, we have conducted a bibliographical review and visited television information centers to determine the requirements, to then capture the necessary changes in the mechanisms of system automation.

Keywords

Emotion recognition / document automation / digital audiovisual documentation / Television files / Biometrics / Video recognition / Audio recognition

Recibido: 20/04/2019
Aceptado: 28/05/2019
DOI: https://dx.doi.org/10.5557/IIMEI10-N18-052068
Descripción propuesta: Caldera-Serrano, Jorge 2019. Automatic recognition of emotions in the description of motion pictures for television archives. *Métodos de Información*, **10**(18), 52-68

1. Media documentation and biometrics

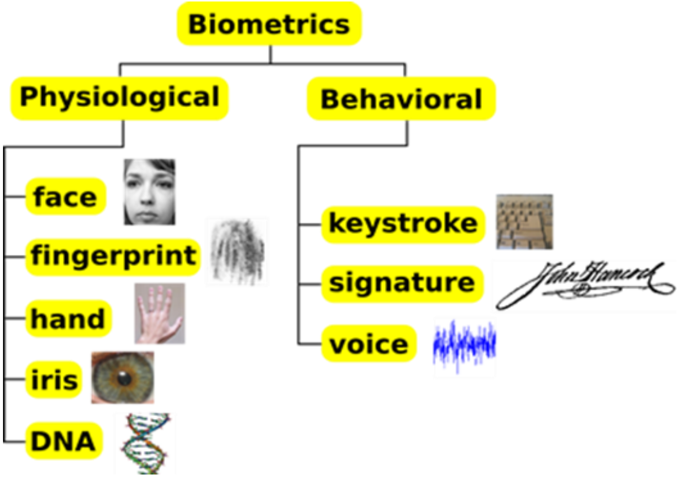
The media are large enterprise schemes where information control is necessary from the perspective of System Theory. The so-called "Fourth Power" holds and controls a large amount of information to be analyzed, guarded and examined with documentary criteria. Such is the workload in the media, particularly in television, that it is necessary to automate as many document management elements as possible (Caldera 2009) (Blanco & Póveda 2015).

Audiovisual information management in television is a complex system and particularly slow in its implementation. Identifying the elements that we can see and hear is not a matter of easy solution in these moments of exploitation and overabundance of material in television archives. Complete dedication to image description is currently impractical, as we need tools and techniques that help us lower the costs of documentary analysis (Servicio 2015).

Biometric technologies are not yet well developed in the management of television archives, however it seems appropriate to reflect on the potential offered by the automatic recognition of images and sound, which will help us identify onomastic and even geographic and thematic visual and sound elements (Caldera 2008).

Biometrics (from the Greek "bios" = life, and "metron" = measurement) is the technology focused on security -and as such, on identification-, which is based on the recognition of the physical, inherent, individual and non-transferable characteristics of the people. It is an automated system that uses the same system as our brain since it recognizes and distinguishes one image from another.

There are different methods of identification and authentication that can be divided into those related to physiology or behavior. The first would include the geometry of the hand, iris, retina, facial recognition and fingerprint as the most studied elements; while those related to behavior include the study of the signature, voice and keyboard dynamics.



Source: http://bio-metrica.com/biometric-theory

This technique accounts for several centuries of history, it is not a novelty or science fiction. Physical parameters were used in Egypt to verify the persons involved in trade. Elements such as birth marks, scars, colors, eyes and teeth were also used in ancient agricultural communities where the materials were stored communally. In Chica for example, palm imprints were stamped to distinguish young children since the fourteenth century.

The nineteenth century is characterized (in biometrics) by the research to identify elements to individualize people, especially for judicial purposes. *Alphonse Bertillon*, head of photography of the Police Department in Paris, developed an anthropometric system consisting in determining the length and width of the head and body parts, as well as the identification of scars and tattoos. But it was not until the twentieth century when the use of fingerprints became a reliable, feasible and safe method of identification. The advancement of technology, especially of information technology derivatives, makes it possible to have other developments such as identity verification through the iris, voice, etc.

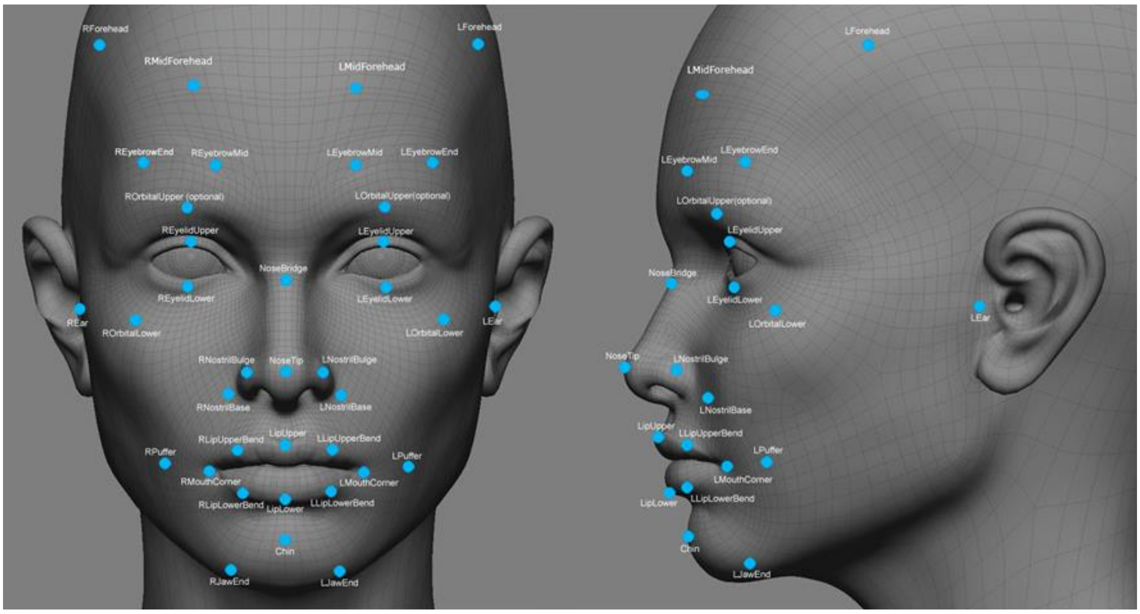
For this reason, biometric identification applies mathematical and statistical methods to identify persons exactly as they are, and not by external elements, such as some sort of credentials. Biometric techniques are based on measuring users to recognize them automatically by applying statistical and Artificial Intelligence techniques (fuzzy logic, neural networks, etc.).

2. Facial and audio recognition

The main development for video recognition is given by the facial recognition studied and developed from biometric studies. Biometric discipline uses biological information to verify identity. The basic idea underlying biometrics is that our bodies contain properties and unique information that can be used to distinguish and individualize us from other individuals in the group. The study and analysis of facial recognition has been conducted from various perspectives, such as Computer Science, Psychology, Forensic Medicine and Anthropometry (Moreno 2004).

In the case of facial analysis, we will see that there are data that set us apart from the rest, both in width, distance between facial elements, such as color and even the existence of wrinkles. These are the elements that structure or shape the characteristics of each face. More than 80 nodal points may be identified for the unequivocal identification of the human face. Normally, the different programs developed do not use all these nodal points but they simply analyze fewer elements thereof (Jain 2004).

These nodal elements are evaluated, creating a numeric code for each element that results in a series of numerical data included in a database. Therefore, the software works with numerical information and not precisely with audiovisual information. The set of data, codes and elements that distinguish each individual is called "faceprint".



Source: [http://endthelie.com/2012/08/24/fbi-sharing-facial-recognition-software-with-police-departments-across-america/.](http://endthelie.com/2012/08/24/fbi-sharing-facial-recognition-software-with-police-departments-across-america/)

There are two stages of implementation for these systems; the first one is system training, in which we introduce the physiognomic data of the individuals by using a sensor, such as cameras or webcams, and store them in a database. The second stage will be the capturing of images and their subsequent comparison (Adler 2007).

There are several criticisms and limitations of facial recognition; for example, preliminary storing for face detection is necessary, and it therefore works with a set of pre-established data. This drawback that is evident and obvious, continues to allow the use of facial recognition and adds real value to it, provided that the conditions observed are optimal, both in capturing the image stored in the database, and also with the one given and compared with later on. However, the limits of good results will arise from elements such as lighting control, angle, distance from the capturing target, the quality thereof, etc., among other things.

Audio recognition

Audio recognition is much more developed than video recognition, surely due to the ease of implementing computer engineering systems. Automatic Speech Recognition (ASR) systems allow access to information through speech, and it is particularly developed in the field of mobile telephony and information technology.

The main problem with ASR systems is that the communication channel is unpredictable, which could lead to reception difficulties due to sound-distortion elements (speaker, microphones, environment, etc.).

There are major advances for the visually impaired, supporting communication through computer peripherals.

Three lines of action are implemented from ASR (Espinosa 1998):

Automatic dictation systems. These are dictation programs with a number of built-in words. Apart from these, they pose serious problems in including new terms.

IVR (Interactive Voice Response) systems. IVR applications are a bridge between people and databases to connect the individuals with computers.

Automatic Speech Segmentation and Labeling systems. These would be automatically labelled phonetic segments. The speech synthesizers used in Text-Voice Conversion require a large number of labeled segments for the synthetic voice to sound natural.

3. Emotions and television audiovisual documentation

The meaning of signs emerges as a representation of reality, mainly, as a subjective experience that serves as a means for social interaction. Since this construction is collective, created by a social group in a more or less established environment, this meaning is therefore a social product.

The correct interpretation of the meaning involves functions that help structure thinking and serve as an instrument of social communication. Hence the meaning of a set of signs (codes) depends on the tacit social agreement to define its use and a consensus on its definition. To interpret the signs, we will not just need to know the signs but also the communicative intention (Eco 1991).

Denotation is defined as the opposite to connotation. The first is the fundamental factor of communication and an essential part of the operation of a language by establishing an associative link for presentation. To denote is to name, to designate. Denotation is part of the technical and scientific discourse. In contrast to denotation, there is connotation. Different signs have affective-emotional nuances. Connotation associates the underlying meanings thereof. As a result, this connotation will vary according to the culture in which the communication takes place, the time period in which it unfolds, the social group, experiences and individual motivations, etc., given that they are unstable and often subjective and individual meanings.

Red indicates a precise color identified in the wavelength, while its connotations are political and even cultural. Just to show, by following the colors, that depending on the culture, the expression of mourning may be related to black or white clothes.

There is a clear difficulty in identifying connoted elements. There is a personal, non-transferable connotation, which is outside the scope of Documentation Sciences for it to be analyzed properly. The reason why certain music reminds us of situations or people is merely related to personal experiences or intimate reflections, not always resulting from reality. This personal connotation is not analyzable and also lacks interest.

What matters is collective connotation, or it is at least valid for the group of documentalists-journalists-viewers of a society at a given time, who have a connoted global view of the same sign.

As always, documentary work is nothing but a reflection of the needs of users, and they do not always require specific images to remedy a lack of information or to use it in a piece of news, but many times what they need is material to contextualize, explain, and even serve as a fill-in information.

In these cases their requests are usually related to connoted elements, because the effort of translating denoted elements to connoted elements is only required to not take place at the time of the consultation but that it is developed at the time of document analysis, and if possible, with automatic techniques derived from biometrics.

In the field of audiovisuals there is a triple reality that must be analyzed: the reality provided by the visual track, the reality identified by the soundtrack and the relationship, not always synchronous, between the aforementioned. The audiovisual character of the material is what undoubtedly enhances connotation, especially those sound aspects.

The main problem in analyzing audiovisual connotation is not so much discriminating between what is useful for a group or for the individual but rather how to designate that emotion or feeling, and that it is understood unequivocally and unambiguously by all parties involved.

Our connoted elements will not precisely be a descriptor but an attribute that can be given to such descriptors, and to which images may also be associated, which in turn will have other descriptors. That is, a connoted attribute, as we propose, will always be associated to one or more descriptors.

A useful list of possible emotions attributes would be:

- Satisfaction: pleasure / disgust / loathing.
- Access to wealth: ostentation / wealth / poverty / misery / famine.

- Attitude: courage / cowardice.
- Love: attraction / desire / tenderness / passion.
- Tiredness: vitality / exhaustion / laziness.
- Anger: bad temper / outrage / exasperation.
- Mood: euphoria / happiness / serenity / sadness / despair.
- Humor: grief / crying / sobbing / smile / laugh / chuckle / hilarity
- Fear: nervousness / alarm / horror / panic / shock
- Surprise: astonishment-wonder / admiration.
- Use of time: boredom / fun.
- Appreciation: admiration / pride / shame / embarrassment.

4. Identifying emotions for television

Much has been analyzed and written about emotions, since the first studies conducted by Darwin. Ekman (2004) states that his interest lies in the fact that emotions are cross-cultural and thereof understandable anywhere in the world. He also points out the six basic emotions: anger, disgust, fear, happiness, sadness and surprise.

The concept of emotion involves, according to Schmidt-Atzert (1985, pp. 35), a subjective experience, a physiological reaction and a behavior. Wundt (1910) classified emotions into pleasant and unpleasant, while Scmidt-Atzert (1985) grouped them into joy, pleasure, affection, friendliness, longing, anxiety, aversion, aggression, sadness, perplexity, envy and fear.

It is not the purpose of this paper to analyze emotions from the psychological point of view, since there is profuse scientific literature on the matter and it is also not our specialty, but we note that such literature accepts that emotions are communicated by both facial expressions, tone of voice and gestures, in some cases.

Emotions are a key determinant in human communication; as noted before, there are several ways of communicating emotions, such as facial expressions, posture, tone of voice, choice of words, respiration, body temperature, etc.

Emotions change the meaning of the message in many cases (Diego Martin 2006).

Biometric studies can only focus on analyzing the recognition of expressions in faces and speech parameters, and it is precisely these parameters that we will analyze and concentrate.

Audio emotions

It has been established that emotional meaning, both acoustic and lexical, is detected in linguistic characteristics. The difficult part certainly lies in parameterizing and defining the characteristics of each of these emotions.
Murray and Arnott (1993) parameterized these speech elements, gathered by Martin de Diego (2006), in the following table:

	Anger	***Happiness***	***Sadness***	***Fear***	***Displeasure***
Speed	Slightly fast	Fast or delayed	Slow-paced	Very fast	Much faster
Variation	Very high	High	Slightly low	Very high	Very low
Range	Wide	Wide	Narrow	Wide	Wide
Breathing	Synchronized	Synchronized	Resounding	Irregular	Grumbling
Intensity	High	High	Low	Normal	Low
Articulation	Tense	Normal	Slow-paced	Accurate	Normal
Voice quality	From chest	Strident	Resounding	Irregular	Rumbling

Source: Martín de Diego, et. al, 2006.

Table 1 classifies the physical characteristics that can be parameterized with emotions, which can be then identified in a previous database, to which the analyzed audiovisual documents can be subsequently compared.

However, there are many phonetic variations that define voice, and as such, the above parameters could confuse us in some cases, or fail to give us exact data but only approximates.

From an information systems point of view, there are already tested studies that indicate that emotions can be detected through voice, such as those conducted Dumouchel (2009) and Grimm (2007).

Emotions in video

The gestures of the face, just like voice, are the best way to recognize and identify emotions. This has been equally tested in still images, where such work can be quite complex, and therefore using them in audiovisual information can be relevant and with better results, given that we have the same character for several keyframes, and emotions by definition are changing over time.

Through biometric techniques we delimit faces in images (identifying oval shapes with the eyes and nose, etc.), so that the system can track the face. Until recently it was necessary to control the lighting and the position of the face. Rapid changes are being developed in biometric studies, and as such, these conditions (though always improving the outcomes) are not mandatory.

Ekman (1972) was the great pioneer in recognizing emotions in facial expressions, identifying the six major emotions. Of the different techniques provided (approaches based on optical flow, tracking of characteristics and approaches based on the alignment of the model), the best results in biometrics are obtained from the comparison of the data included a priori in a database, numerical data resulting from the previous indication of the physical characteristics of each emotions. As easy, or difficult, as including still and moving image material into the system, which point out the predominant emotion.

There are multiple analysis techniques, as stated by Martin de Diego (2006), that allow the technical feasibility of carrying out such work. More specifically, Castrillón (2008) offers interesting work on the different techniques for the extraction of facial characteristics, specifically facial expressions.

Any biometric analysis tool must first be taught through training and further instruction in order to conduct the audio and visual analysis.

First, we must present a sample collection to provide the machine with the numerical parameters corresponding to each of the emotions, so that the machine can really compare these data with future results. It is a fundamental stage for attaining a correct result.

Once the system has been trained, we must confirm its correct operation using the emotions recognition test.

Neural networks, support vector machines, classification trees and k-neighbors, are some of the techniques used, and the first to obtain the best results.

5. Optimizing the system to identify emotions in TV

Martin de Diego (2006) says that we can get better results by joining the characteristics of audio and video emotions during the training of the system, as well as for the identification in a real environment. Although his studies are conducted with static images, the results of facial identification in moving images using biometric techniques are encouraging, so what we will have to do is incorporate emotions parameters into the system so that, while the person is being identified, the emotion in the face and/or the voice of the character can be also be detected.

Therefore, based on the problems arising from the use of a single development, it seems correct to apply both techniques separately (emotion recognition through audio and video) so as to compare results, with the system assigning value percentages of the emotion through hearing and image, as well as verification through the interaction of the two components. In the meantime, and due to the lack of infallibility of biometric systems, validation by the human component will continue to be required.

Analyzing emotions jointly will narrow possible errors derived from hearing, mainly.

Another recognized advantage is the use of moving images. Virtually all studies conducted point out the problems in identifying emotions from static images. In the case of audiovisual material, what we really have are photos overlapping quickly, giving the impression of movement. From the point of view of recognizing movement, the proposal is to have characters in different keyframes, maintaining or altering their expression, and also the position of the head. This facilitates the identification and analysis of the emotions of the character at different times. And evidently, there is no problem of confusion between characters since biometrics features analysis techniques for contours, edges, etc. to identify each of the characters or at least to detect any changes thereof.

Therefore, automation in the identification of emotions will be valid if biometric identification is being applied to refer to the characters. And as such, it not only serves to identify emotions, but also specific characters associated with these emotions.

Moreover, emotions may be identified in anonymous characters, which may be required at some point by journalists or the users of these documentation services.

For this reason the work of documentalists remains a double function:
Data validation and recovery ultimately define up to what point are the obtained results correct. This work must still be maintained due to the potential errors of biometric recognizers.

The second section relates to the maintenance of the biometric system. We must train the biometric devices identifying emotions, and we must constantly incorporate new characters that start to become news or to be required by journalists.

6. Conclusions

Biometrics and information centers in television networks are bound to meet. Especially in digital work environments in which the reading of information can and should be automated.

We must improve human document management due to the high costs in human resources, and therefore the economic implications involved.

Biometrics not only facilitates the option of recognizing persons, but also of associating the emotions reflected in images and audio to them, thus enhancing their future use and specifying the validity of the recovered items.

From the information systems perspective this is already possible as there are efficient tools and technologies for this purpose. We only need to create large image databases that gather all these television characters (who are not as many as we think) and identify their emotions, which has been proposed in this paper.

Journalists require this, the technological conditions exist (with increasing quality), and it is therefore the duty of the documentation departments to start working on system automation lines to deliver better results and better products to our users.

Acknowledgements

This work has been funded by the Board of Extremadura (Office of Education, Science and Technology) and the European Social Fund under the support plan for the activities conducted by the Research Groups registered in the catalogue of the Board of Extremadura. GR10019.

7. Bibliografía

ADLER, A. y SCHUCKER, M., 2007. Comparing human and automatic FACE recognition performance. *IEEE Transactions on Systems Man and Cybernetics part B-Cybernetics* **37**, 1248-1255. ISSN 1083-4419. DOI: 10.1109/TSMCB.2007.907036

BLANCO-IZQUIERDO, M.J. y PÓVEDA-LÓPEZ, I.C., 2015. Metadatos documentales en Canal Extremadura Televisión. *Cuadernos de Documentación Multimedia*, **26**, 13-132. ISSN 1575-9733. DOI 10.5209/rev_CDMU.2015.v26.50633.

CALDERA-SERRANO, J., 2006. Terminological control of "anonymous groups" for catalogues of audiovisual television. *Journal of Librarianship and Information Science*, **38**(3), 187-196. ISSN 0961-0006. DOI 10.1177/0961000606066582

CALDERA-SERRANO, J., 2010. Group connotation in the analysis of the imagen in motion used in television department. *Journal of Librarianship and Information Science*, **42**(1), 70-75. ISSN 1575-9733. DOI 10.1177/0961000609351375

CALDERA-SERRANO, J. y ZAPICO-ALONSO F.. 2009. Biometric facial identification. *El Profesional de la Información*, **18**(4), 427-431.ISSN 1699-2407 DOI 10.3145/epi.2009.jul.11

CASTRILLON, W.A.; ALVAREZ, D.A. Y LOPEZ, A.F., 2008. Técnicas de extracción de características en imágenes para el reconocimiento de expresiones faciales. *Scientia et Tecnica* **38**, 7-12. ISSN 0122- 1701. DOI 10.22517/23447214.3681

DUMOUCHEL, P. [et al.], 2009. Cepstral and Long-Term Features for Emotion Recognition. En: *Interspeech* 2009, 344-347. Brighton, UK. Communication Association. https://pdfs.semanticscholar.org/6d9f/d594c357bbc110e80398af7b853a5e8cfb70.pdf?_ga=2.72273363.1988264968.1563431601-1549472693.1563431601

ECO, H., 1991. *Tratado de semiótica general.* Barcelona: Lumen. ISBN 9788426411228

EKMAN, P. (ed.), 1972. *Emotion in the Human Face.* New York: Cambridge University Press. ISBN 9781483147635

EKMEN, P., 2004. *Emotions Revealed.* New York: Times Books. ISBN 0-8050-7275-6

GRIMM, M. [et al.], 2007. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, **49**(10-11), 787-800. ISBN 0167-6393. DOI. 10.1016/j.specom.2007.01.010

HUBERMAN, A.M. y MILES, M.B., 1994. Data Management and analysis methods. En DENZIN, N.K. and LINCOLN, Y.S., (Eds.). *Handbook of Qualitative Research*, pp 428-444. Los Angeles, California: Sage Thousand Oaks.

JAIN, A.K.; ROSS, A. y PRABHKAR S., 2004. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, **14**(1), 4-20. DOI: 10.1109/TCSVT.2003.818349

MARTIN DE DIEGO, I.; SERRANO, A.; CONDE, C. y CABELLO E., 2006. Técnicas de reconocimiento automático de emociones. *Revista electrónica Teoría de la Educación: Educación y Cultura en la sociedad de la información* **7**(2), 107-127. ISSN 1138-9737.

MORENO DIAZ, A.B., 2004. *Reconocimiento facial automático mediante técnicas de visión tridimensional. Madrid:* Universidad Politécnica. http://oa.upm.es/625/01/10200408.pdf . Tesis doctoral.

MURRAY, I.R. y ARNOTT, J.L., 1993. Toward the simulation of emotion in synthetic speech: A review of literature on human vocal emotion. *Journal of Acoustical Society of America*, **93**(2), 1097-1108. ISSN: 0001-4966. DOI 10.1121/1.405558

SCHMIDT-ATZERT, L., 1985. *Psicología de las emociones.* Barcelona: Editorial Herder. ISBN: 8425414539

SERVICIO DOCUMENTACIÓN MULTIMEDIA, 2015. Multimedia y nuevas tendencias tecnológicas en documentación informativa a propósito de Bibliored 3.0. *Cuadernos de Documentación Multimedia*, **26**, 145-152. ISSN: 1575-9733. DOI 10.5209/rev_CDMU.v26.50635

WUNDT, W., 1910. *Grundzüge der physiologischen Psychologie*. Leipzig: Engelmann.