EPiC Series in Computing

Volume 69, 2020, Pages 256–263

Proceedings of 35th International Conference on Computers and Their Applications

Smart Ecosystems through Voice and Images

Alexander Iliev[1] and Peter L. Stanchev [1,2]

[1] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

[2] Kettering University, Flint, USA

ailiev@berkeley.edu, pstanche@kettering.edu

Abstract

In this article we summarize at a high-level some of the popular smart technologies that may contrive many smart city ecosystems. More specifically we will emphasize the automation of various processes based on the extraction and analysis of digital media, through speech signals and images. Currently, there are many productized systems for personalization and recommendation of digital media content as well as various services in different areas. Most of them are developed with human-machine interaction in mind. Usually, this is done through a conventional use of a mouse and a keyboard. The user types their response manually, which is then recorded by the system for further analysis.

1 Introduction

Many factors may be able to contribute to the development of high-level smart concepts. On the one hand, the advancements of handheld devices inevitably create the basis for the realization of many smart technologies. This is due to the increase of computational power and minimization of electronics worldwide. On the other hand, the exponential increase of data according to different sources leads to a growing necessity for the development of adequate technologies that can clean, structure, and analyze more and more data every day. Different kinds of recorded media such as video, images, music, or speech can successfully be used as a source of automation in number of smart services. Considering speech as a source for automation is not only an advanced way to think about future emerging technologies, but instead, it is already a necessity. Due to the advancement of sophisticated technologies and the growth of data, the marriage between Digital Signal Processing (DSP) and Artificial Intelligence is inevitable and is already happening. After carefully extracting, cleaning, and organizing multiple features from speech, we can devise sophisticated feature vectors containing attributes, collected through DSP, then passed on to machine learning algorithms to complete specific smart tasks. Speech carries a vast amount of information on many levels. Through speech we can understand: a) *who is talking*, b) *what they are saying*, c) *what is the gender of each speaker*, d) *how speakers are talking using accents and dialects*, e) *determine the origin of the speakers*, f) *determine the emotional state they are in*, and last but not least g) *deduce numerous health-related issues in some cases*. All of these can

be contributing to layers of information when forming an automated decision in the development of smart technologies.

2 Smart Systems using Voice

Creating smart personalized systems is not only possible through textual information that we typed. There is another, more sophisticated level of Smart System development, especially when it comes to personalization. One of the main parameters of such systems is the way we gather the specifics of human activity or behavior in a natural environment of our everyday activity. This means that the systems gathering this vital source as a main vehicle for creating our smart logic must remain stealthy and non-intrusive. The criteria for all of these are easily met through one of the most natural ways of human communication – speech. Voice is one of the main vehicles for delivering valuable information when it comes to human activity. It has many benefits such as: a) *the way it is collected*; b) *the type of processing applied to it* or c) *the layer of information gathered through voice activity*.

As for the first benefit, voice is one of the easiest non-intrusive sources of information we can collect. Recording through microphones is both cheap and stealthy. Regarding the second benefit, processing voice signals is trivial, as it does not require the level of computational complexity video and image processing does. These two benefits make working with speech extremely attractive when it comes to developing speech based smart systems. The third important benefit when using speech for smart system development is that information can flow in many different layers.

2.1 Smart Services in the Medical Field

Smart Systems through voice have been applied to number of fields already. One of the most beneficial fields is the medical field. There are number of services that can directly benefit from the advancement of this technology. When combining Digital Signal Processing with Deep Learning great products are already implemented. One such product is monitoring of patients as described in [1]. This paper suggests a method that implements a quick reaction time of medical personal when dealing with disabled people. More specifically, the method uses a Raspberry Pi (RPi) based on Internet of Things (IoT) device for patients with disabilities. The device is activated via voice and is interpreting voices using a Support Vector Machine (SVM) then generating alerts for medical personal. This research was directly applied in real-world hospital scenarios. RPi devices have the advantage of being relatively cheap for the functionality that they offer. They are also small in size and very lightweight for the processing and implementation power that they offer for a number of IoT applications. The board comes with a 1 GB memory and 1.2 GHz Quad core processing power [2]. Scripts written in Python have been used to enable and control the voice recording part of the system as well as the communication part via liquid-crystal display (LCD) touchscreen [3]. Raspberry Pi boards use a mini version of Linux called Raspbian and is running from a micro SD card. Based on the card the user has different access times and speed respectively. In this sense the system boots from the card onto the RPi device just as a plug and play device.

The Fourier Transform is usually used when extracting and collecting the features. In this case, the Discreet Cosine Transform (DCT) was used to obtain the features [4] as shown in Figure 1.

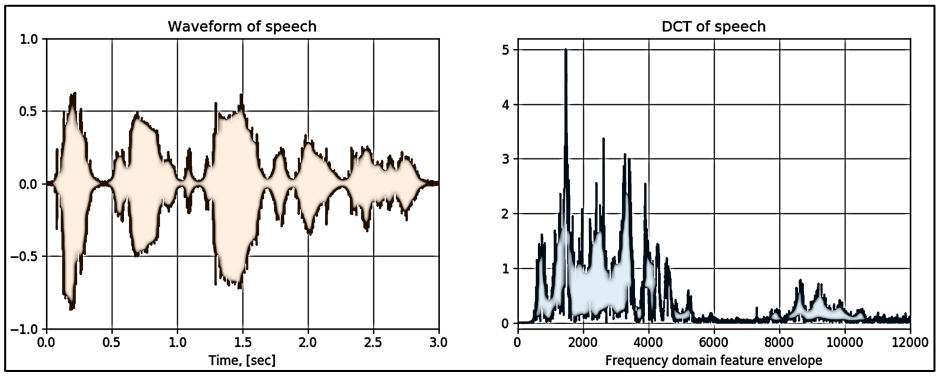

Figure 1: Original voice waveform and its DCT feature frequencies

As can be seen in Figure 1, this multiclass machine-learning algorithm used Support Vector Machine (SVM) as part of its core classification feeding voice waveform parameters into the input. The signal is fed into a DCT block, and then frequency features from DCT domain are fed into the SVM for classification. The dataset used was the Montreal Affective Voices (MAV) [5]. It contained a diverse set of voice samples containing emotions for both training and testing of the classifier. More specifically, the following nine emotional states were represented in 90 non-verbal voice samples: *happiness*, *anger*, *sadness*, *surprise*, *disgust*, *fear*, *pain*, *pleasure*, and *neutral*.

2.2 Smart Systems in Entertainment using Emotions

New movies, TV series and music albums are being created every day. This alone makes the task of finding and recommending content extremely daunting. The solution to this kind of problem can be through automated background recordings of user behavior with a one-time consent of the user. In this way we can create systems that are very powerful, with natural human-machine interaction.

There is some research that is pointing in this direction already as shown in [6]. In this work we suggest a service-oriented architecture (SOA) for access into digital resource libraries through voice. Verbal communication between users and computers has been looked at from another prospective. We use emotion recognition from speech signals as a main vehicle for convening specific needs, which can be applied directly to content recommendation. For the task, some specific speech parameters and techniques were established previously in [7]. There, six different emotional states were adopted: angry, happy, neutral, sad, fear and surprise. Some of the most important classical prosodic features in time and frequency domain were also discussed.

This work was further extended for the more practical case in recommendation system development [8], where gender separation from voice also plays a role in user separation. It is especially important when there are more users communicating at the same time either amongst them or with the entertainment system. One of the most prominent features in emotion recognition from voice was proven to be the Glottal Symmetry (GS) as described in [9, 10] and further studied in [7]. This feature domain has a very specific inverse filtering technique, but the payback is great as it provides a very robust domain for recognizing emotions from speech.

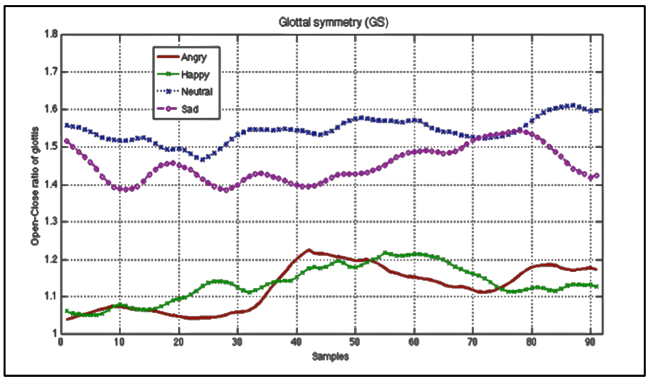

Figure 2. Glottal symmetry for four emotional states [11]

In Figure 2 [11], a plot of the Glottal Symmetry for four emotions is given. For the most part, we can clearly distinguish among different emotional states based on voice and this feature alone. The idea to use GS as a feature vector sparked the interest of the testing of new machine learning classifiers and it was one of the first feature domain tested with an Optimum-Path Forest classifiers as described in [12]. More detailed summary of other conventional machine learning techniques and various features is given in [13].

After all of these features were tested [6-13], there was an effort to implement a more productized version as described in [14], where sentiment from five different books from "The Game of Thrones" was extracted using Natural Language Processing (NLP) techniques implemented using the Natural Language Processing Toolkit (NLTK) in Python. The results from finding the sentiments from these five books were then mapped to the results from speech, where emotions were detected using the features and techniques as described before. As a result, a novel content discovery and selection system was designed. It was solely based on human behavior. The conclusion from this work was that machines could learn the behavioral patterns from multiple speakers and consequently could recommend content to them. This same idea can be applied to electronic books, TV series and digital movies libraries, as well as music players as described next.

3 Smart Systems through Images

Images are a great source of information. Every second, people all over the world create millions of images. The widely available image-capture techniques, inexpensive storage, and dissemination methods have made videos and images convenient and easily available. This in turn increases the availability for solutions to the problems in indexing and retrieval.

While major search engines are in the process of rolling out audiovisual search capabilities, such descriptions are definitely not sufficient. Context is important in these scenarios and must be managed to make such searches truly useful. In light of these issues, researchers around the world have begun to work on multimedia semantics to study the measured interactions between users and real objects, with the ultimate goal of trying to satisfy the user community by providing them with the objects they require.

3.1 Finding Similarity in Images

A very important direction towards the support of content-based image retrieval is feature based similarity access using high-level semantic features based on the extraction of low-level color, shape

and texture characteristics and their conversion into high-level semantic features using fuzzy production rules, derived with the help of an image mining technique [15].

Image recognition is used in the search of products or objects, by firstly identifying the objects, and then searching the network for similar patterns as outlined in [16].

Figure 3: Claude Monet Japanese Bridge paining and a personal photo of the same bridge

How to find "if the two bridges are the same bridge" is discussed in [17]. A comparison is performed between Claude Monet "The Japanese Bridge (1899)" (https://www.wikiart.org/en/claude-monet/the-japanese-bridge, public domain as seen in January 2020) and a personally taken photo by the author on the same bridge – Figure 3.

In [18] retrieval by contrast is presented. Examples for search by different contrast are for instance: light-dark contrast in "Dr. Faustus in his study room" sketching by Rembrandt and "Guitar on Mantelpiece" by Paulo Picasso; cold-warm-contrast in "Le Moulin de la Galette" by Auguste Renoir, and "Houses of Parliament" by Claude Monet; simultaneous contrast in "Stripping of Christ" by El Greco, and "Café at Evening" by Vincent van Gogh.

3.2 Smart Services for Medical Images

Machine learning, with regard to deep learning, helps to identify, classify, and quantify patterns in medical images, which are described in [19]. In the medical multimodality and multimedia systems the main effort is put on advancements in technology that have increased the capability to produce images, to manipulate them and improve the medical diagnosis.

In [20] the retinal vessel radius estimation is studied and a segmentation method for vessel centerlines based on ridge descriptors is processed. The study on radius estimation reveals that the radius estimation by the matched filters based on the second order derivatives of Gaussian kernels is only correct at the vessel center. The relation between the vessel radius and the scale of the Gaussian kernel in the estimation method based on the normalized largest curvature is also studied. The ridge descriptor proposed contains the normalized largest curvature and the orientations of gradients in the local neighborhood. For vessels of a certain scale, the distribution of the descriptors is assumed to be a normal distribution and is learned from a training set with known truth. Vessel centerline segmentation can be then performed based on the distance between the ridge descriptor at candidate pixels and the learned model.

3.3 Extracting Higher-level Visual Features

A tool for extracting higher-level visual features for art painting classification based on MPEG-7 descriptors was implemented in the system "Art Painting Image Color Aesthetics and Semantics" [21]. The approach consists of the following steps: (1) tiling images into non-overlapping rectangles in order to capture more detailed local information; (2) the tiles of the images are clustered for each MPEG-7 descriptor; (3) vector quantization is used to assign a unique value to each tile, which corresponds to

the number of the cluster where the tile belongs, in order to reduce the dimensionality of the data. The distribution of significance of the attributes, the importance of the underlying MPEG-7 descriptors as well as analysis of spatial granularity for class prediction in this domain is analyzed.

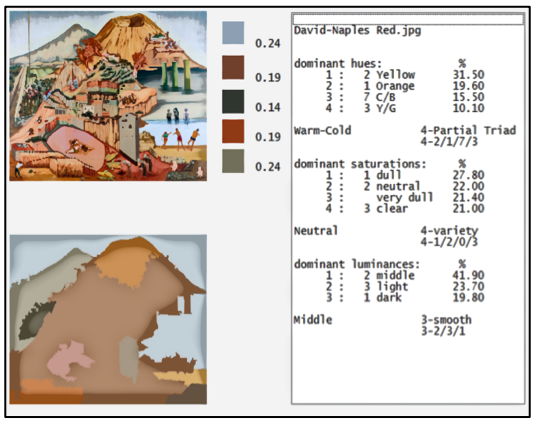

Figure 4. Harmonies attributes for Nurit David's artwork "Naples Red"

Figure 4 shows the tiles, closest to the centroid of Color Structure Descriptor. The idea of this presentation is that these tiles can be used later as elements in a visual lexicon for representing specifics of some image profiles.

4 Conclusions

As we have shown, there are a number of areas that already use smart systems at their core for improving of services in healthcare, entertainment, robotics, mobility, online learning, etc. Many more systems can also benefit from this kind of technological advancements such as security, banking and various services in medicine or even warfare. The two main vehicles to deliver such sophisticated systems are voice and images, but more widely used features from textual cues can also be used. The two domains used herein show how through vision and sounds many diverse services can be improved in our everyday life. Further studies in these and other areas can further rally and fine-tune all **s**ystems aiming to advance any smart system technology included in the global IoT ecosystem.

References

[1] G. Latif, A. Khan, M. Butt, O. Butt, “IoT based Real-time Voice Analysis and Smart Monitoring System for Disabled People”, Asia Pacific Journal of Contemporary Education and Communication Technology (APIAR), Vol. 3, Issue 2, pp. 227-234, DOI:10.25275/apjcectv3i2ict5, ISBN (eBook) 978 0 9943656 8 2 | ISSN: 2205-6181, 2017

[2] B. Upton and G. Haloacene, “Raspberry Pi user guide”, John Wiley & Sons, 2014

[3] B. Smith, “Raspberry Pi Assembly Language RASPBIAN Beginners: Hands on Guide”, CreateSpace Independent Publishing Platform, 2013

[4] S. Kumar and T. RangaBabu, “Emotion and Gender Recognition of Speech Signals Using SVM”, Emotion, 4(3), 2015

[5] P. Belin, S. Fillion-Bilodeau and F. Gosselin, "The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing". Behavior research methods, 40(2), pp.531-539, 2008

[6] A. Iliev and P. Stanchev, "Smart multifunctional digital content ecosystem using emotion analysis of voice", 18th International Conference on Computer Systems and Technologies CompSysTech'17, Ruse, Bulgaria – June.22-24.2017, ACM, ISBN 978-1-4503-5234-5, Volume 1369, pp.58-64, ISBN: 978-1-4503-5234-5

[7] A. Iliev, Monograph: "Emotion Recognition from Speech", Lambert Academic Publishing, 2012

[8] A. Iliev and P. Stanchev, "Information Retrieval and Recommendation Using Emotion from Speech Signal", in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval, Miami, FL, USA, pp. 222-225, DOI:10.1109/MIPR.2018.00054, April 10-12, 2018

[9] A. Iliev and M. Scordilis, "Spoken Emotion Recognition Using Glottal Symmetry", EURASIP Journal on Advances in Signal Processing, Volume 2011, Article ID 624575, ISSN: 1687-6180

[10] A. Iliev, M. Scordilis, "Emotion Recognition in Speech using Inter-Sentence Glottal Statistics", proceedings of the 15th International Conference on systems, Signals and Image Processing, IEEE-IWSSIP 2008, June 25-28, 2008, pp. 465-468, Bratislava, Slovakia

[11] A. Iliev and P. Stanchev, "Glottal Attributes Extracted from Speech with Application to Emotion Driven Smart Systems", in proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 1: KDIR, pages 297-302 ISBN: 978-989-758-330-8, indexed in Thomson Reuters, Seville, Spain, 18-20 September 2018

[12] A. Iliev, M. Scordilis, J. Papa, A. Falcão "Spoken emotion recognition through optimum-path forest classification using glottal features", Journal on Computer Speech and Language, ELSEVIER, Vol. 24, Issue 3, 2010, pp. 445-460, ISSN: 0885-2308

[13] A. Iliev, "Feature vectors for emotion recognition in speech", National Informatics Conference, Sofia, Bulgaria, pp. 225-238, 2016

[14] A. Iliev, "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, pp. 233-238, ISBN: 978-1-4503-5622-0, DOI: 10.1145/3281375.3281399, 2018

[15] P. Stanchev, "Using Image Mining for Image Retrieval", IASTED International Conference Computer Science and Technology, May 19-21, 2003, Cancun, Mexico, 214-218

[16] V. Windson, Using images to extend smart object discovery in an Internet of Things scenario, file:///C:/Users/pstan/Desktop/4057-829-4030-1-10-20181009.pdf

[17] P. Stanchev, D. Green Jr., B. Dimitrov, "Some Issues in the Art Image Database Systems", Journal of Digital Information Management, Volume 4, Issue 4, December, 2006, 227-232

[18] P. Stanchev, D. Green and B. Dimitrov, "High level color similarity retrieval", International Journal Information Theories & Applications, Volume 10, Number 3, 283-287, 2003

[19] S. Dinggang, W. Guorong, and S. Heung-Il, "Deep learning in medical image analysis", Annual Review of Biomedical Engineering, Vol. 19:221-248 (Volume publication date June 2017) https://doi.org/10.1146/annurev-bioeng-071516-044442

[20] C. Wu, J. Derwent, P. Stanchev, "Retinal Radius Estimation and a Vessel Center Line Segmentation Method Based on Ridge Description", J. Sign Process System, 2008

[21] K. Ivanova, I. Mitov, P. Stanchev, E. Velikova, K. Vanhoof, B. Depaire, R. Kannan, "Local Features in APICAS (Analyzing of Added Value of the Descriptors Based on MPEG-7 Vector Quantization)", Int. J of Computer Science and Artificial Intelligence, 2(4), Dec. 2012, pp. 23-32, ISSN: 2226-4450 (online) 2226-4469 (print)