Smart Multifunctional Digital Content Ecosystem Using Emotion Analysis of Voice

Alexander I. Iliev, Peter Stanchev

Abstract: *In an attempt to establish an improved service-oriented architecture (SOA) for interoperable and customizable access of digital cultural resources an automatic deterministic technique can potentially lead to the improvement of searching, recommending and personalizing of content. Such technique can be developed in many ways using different means for data search and analysis. This paper focuses on the use of voice and emotion recognition in speech as a main vehicle for delivering an alternative way to develop novel solutions for integrating the loosely connected components that exchange information based on a common data model. The parameters used to construct the feature vectors for analysis carried pitch, temporal and duration information. They were compared to the glottal symmetry extracted from the speech source using inverse filtering. A comparison to their first derivatives was also a subject of investigation in this paper. The speech source was a 100-minute long theatrical play containing four male speakers and was recorder at 8kHz with 16-bit sample resolution. Four emotional states were targeted namely: happy, angry, fear, and neutral. Classification was performed using k-Nearest Neighbor method. Training and testing experiments were performed in three scenarios: 60/40, 70/30 and 80/20 minutes respectively. A close comparison of each feature and its rate of change show that the time-domain features perform better while using lesser computational strain than their first derivative counterparts. Furthermore, a correct recognition rate was achieved of up 95% using the chosen features.*

Keywords: *Emotion, Voice, Emotion Recognition, Smart systems, Digital Culture Ecosystem*

INTRODUCTION

Using emotion in voice can be of much help when developing novel methods for content search and access via verbal communication between man and machine. In particular such methodologies can be made more robust and practical in everyday life by slimming down the number of emotional states. In this case computational complexity will decrease and confidence levels of the results may lead to more robust analysis. In particular emotional states: *happy*, *angry*, *fear* and *neutral* were used as primary emotional states in this paper. The motivation behind choosing these four emotions is based on the "big six" set [10, 11] as their practical use in real-life environment is well established [12]. There is no specific emotional set that can be used for benchmarking, since the literature shows number of research motivated by different underlying reasoning [1], [2], [3]. The number of speech databases used for research on the topic is also vast [4]. Each database has a specific set of emotions, speaker type in terms of gender, age and life experience (actor or non-actor). In the particular case a mixed emotional set for testing and training a practical system can be designed by using non-actors of different age and gender, hence emulating better the real-world environment.

THE IMPORTANCE OF EMOTIONS IN VERBAL COMMUNICATION

Speech parameters greatly vary by their nature, but generally they are portrayed by swift and abrupt changes and are very typical for each emotion. In a parameter space such features must be chosen so that they have little overlap in order to improve correct detection [5], [6]. One of the most effective feature domains is Glottal Symmetry (GS)

CompSysTech'17, June 23–24, 2017, Ruse, Bulgaria
 ISBN 978-1-4503-5234-5/17/06...$15.00
https://doi.org/10.1145/3134302.3134342

defined as the ratio between opening and closing of the epiglottis in a production of voiced speech [6]. The overlap of different classes representing emotions, using GS was previously established in our works [6, 7] and we are now extending this research with application to the current practical case. By means of Principal Component Analysis applied to six emotional cases (*angry*, *happy*, *neutral*, *sad*, *fear* and *surprise*) the distribution of 1st to 2nd glottal opening to closing ratio was founded. The Glottal Symmetry feature samples for each emotional state form an overlapping cluster at a center point, which signifies that features overlap somewhat. It was also established that a successful recognition rate grows higher when we move away from that point, because of a higher separation of each class in the feature domain. This can be observed as rays coming out of the overlap region; hence each emotion forms a unique cluster outside the center. Another interesting remark that can be made from this research is validating the reasoning behind choosing the correct emotions applicable in our case. In particular, the overall model for description and interchange of data will enable the combination of distributed and heterogeneous multimedia resources including text, images, data, video, audio, etc. It will be based on established standards for formats of data, metadata and interchange of digital objects, connected to emotional content. Based on a particular emotion, previously classified digital cultural content can be sorted according an emotional description. Different methodologies in the recognition task provide various success rates in emotion recognition.

Time-domain Features		
Pitch elements – features & their 1st derivatives		
Mean	[meanP]	[dmeanP]
Median	[medP]	[dmedP]
Standard deviation	[stdP]	[dstdP]
Maximum	[maxP]	[dmaxP]
Rising-falling	[rifaP]	[drifoP]
Max falling range	[maxfrP]	[dmaxfrP]
Temporal Energy vs. their 1st derivatives		
Mean	[meanE]	[dmeanE]
Standard deviation	[stdE]	[dstdE]
Maximum	[maxE]	[dmaxE]
Duration features vs. their 1st derivatives		
Zero crossing rate	[zcRate]	[dzcRate]
Speaking rate	[spRate]	[dspRate]

Table 1: Time-domain features [8].

To further confirm if the chosen four emotional states can be beneficial for our case study, a different set of features was chosen to study their performance. Instead of glottal symmetries, the additional feature vectors used for training and testing had fixed lengths of 11 elements: 6 pitch, 3 temporal energy and 2 durational features as shown in Table 1 [8].

The features used herein can be expressed as:

$$W_F = [F_p \; F_e \; F_d] \; , \qquad (1)$$

and the rate of change is depicted as:

$$W_F' = \left[\frac{dF_p \; dF_e \; dF_d}{dt}\right] , \qquad (2)$$

where,

F_p = (*meanP, medP, stdP, maxP, rifaP, maxfrP*) ,
F_e = (*meanE, stdE, maxE*) ,
F_d = (*zcRate, spRate*)

dF_p , dF_e , dF_d - depict the rate of change in the time domain, and W_F signifies the cumulative feature vector for all the features from the time-domain, F_p represents the pitch atrubutes, F_e depict the temporal energy attributes and F_d signify the feature vector encompassing the two durational features. F_p, F_e and F_d were obtained from the 100 min speech database based on a four emotion set HAFN. There were formed as a sum up or an average of all voiced segments in every statement. The derivatives W_F' of the cross statement attributes show the change of each attribute amid neighboring utterances in the time-domain. Because all features had very different values, which was due to their various origins and in order to avoid divisions by zero, a simple DC-shift was applied to normalize all cumulative feature vectors W_F and W_F'. The latter presented a linear adjustment for each class.

SPEECH CORPUS

The speech database containing speech samples was obtained from a theatrical play that comprised of four male speakers. The audio signal was 100 minutes long and was sampled at resolutions 8kHz and 16-bit. The signal transcription focused around the 4 distinctive emotional classes of interest to this study and all others were not taken into consideration. Naturally, each of the emotional classes was depicted by a distinctive number of spoken events shown in Table 2.

Emotional spoken events	
Emotion:	Total # of UTs:
Happy	303
Angry	403
Fear	131
Neutral	1,215
All	2,052

Table 2. Emotions and the number of corresponding spoken events.

It can be observed from table 2 that the Neutral class (when it contains no emotion) has the largest number of occurrences. It contained 2,052 spoken emotional events not

Number of spoken emotional events		
Emotion:	Average utterance length in [sec]:	Average # of voiced segments per utterance:
Happy	9.22	4.53
Angry	4.81	4.02
Fear	3.91	2.71
Neutral	15.29	3.67
All	8.31	3.73

Table 3. Number of spoken emotional events.

including silence, pauses in the speech, noise, sighs, coughs, etc. One spoken event in this study signifies an utterance, which can be short or long. It can also comprise of one word with various length. An instant of the mean length for every spoken emotion occurrence is displayed in Table 3.

From Tables 2 and 3 can be seen that the number of voiced segments used in this speaker emotion database was 7,654. However, some of them were too short to be

included in the study and therefore were ignored. As a result, *fear* was only represented in 122 voiced segments. To create our training / testing schema using even number of samples from all emotion classes, each class was truncated to 120 spoken events. The adjustment from 122 was because of using the rate of change in order to find the cross utterance derivatives for each attribute. Hence, the number of training / testing spoken events was 480. Knowing this and taking into consideration expressions (1) and (2), each emotion class used here can therefore be expressed as a matrix for each case as shown:

$$W_F^{HAFN} = \begin{bmatrix} F_{p1} & F_{e1} & F_{d1} \\ \vdots & \vdots & \vdots \\ F_{p120} & F_{e120} & F_{d120} \end{bmatrix}, \qquad (3)$$

and

$$W_F^{HAFN\prime} = \begin{bmatrix} dF_{p1} & dF_{e1} & dF_{d1} \\ \vdots & \vdots & \vdots \\ dF_{p120} & dF_{e120} & dF_{d120} \end{bmatrix}, \qquad (4)$$

Further examination of the attribute vectors for each class of emotions, revealed that there were distinguishable differences in each of the four emotion models. An example is the first portion of every feature vector represented by F_p, and specifically for the emotion cases for N (*neutral*) and F (*fear*). As a contrast, when observing the rate of change of the same feature elements from each class it became apparent that the feature models were not as unique and separable as they were when formed by the actual parametric features. It would be fair to argue that there may be more similarity between H (*happy*) and A (*angry*) where a close proximity of the average values of each attribute is observed in Figure 1.

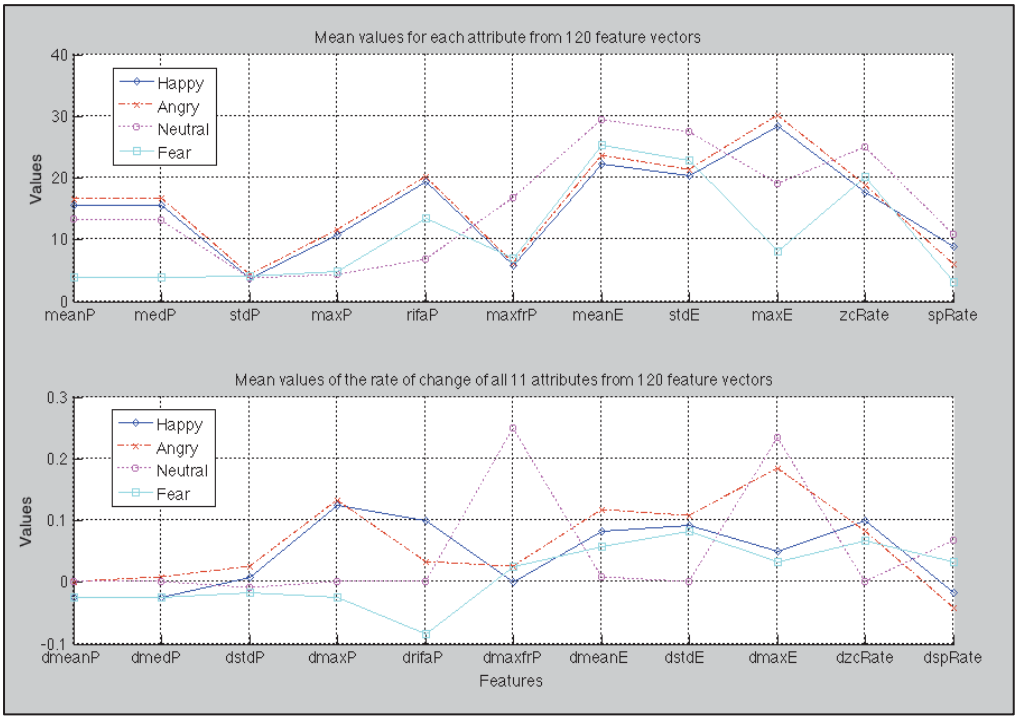

Figure 1: Average feature values and their respective mean rates of change.

In Figure 1 the average value for each feature is depicted. The graph above shows the average values of each attribute from all parametric features while the lower one

depicts the derivatives of the same corresponding features cross utterance in time. Observing the values on the y-axis, we see that change of each attribute as displayed on the lower plot is minimal.

CLASSIFICATION

Because usually there are numerous voiced regions in a spoken event, a holistic view of the entire emotional communication from any given utterance was employed, rather than using a temporal emotion model pertaining to a particular voiced section of an utterance. This notion was established after testing voiced regions chosen randomly across any given spoken event. This is why all attribute vectors were designed based on the mean from the voiced sections of each utterance. It follows that every emotion had to be collected from each utterance as a whole and was viewed globally for each spoken event as defined here. The classification method of choice in this study was the k-Nearest Neighbor or k-NN classifier, which used one nearest neighbor and was implemented by Weka [9]. Several data sets were obtained for training and testing in the classification process.

K nearest neighbor is a simple non-parametric technique that classifies classes by means of measuring their similarities while using a distance function. The latter can vary and can be Euclidean, Manhattan, Minkowski, Chebyshev, etc. Euclidean distance is the most popular measure used and is determined by the expression:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}\ , \qquad (5)$$

where,

K is the number of nearest neighbors to be considered;

x and *y* are the query point and a sample from the examples set, respectively. We choose Euclidean distance.

Any given sample is classified by taking the distance measurements to its neighbors. A sample is assigned to this class that appears to be the most common or closest amongst its K nearest neighbors. Since the distance depends on K, when K=1 the sample is assigned to its nearest neighbor. Once K is selected, predictions based on k-NN examples can begin. When using regression, the prediction based on k-NN is the average of the nearest neighbors specified by K. This is expressed as:

$$Y = \frac{1}{K}\sum_{i=1}^{k} y_i, \qquad (6)$$

where,

y_i represents the i^{th} sample from the examples set and Y is the final prediction result of the query point.

The results from the experimental data after being tested with the k-NN classifier are shown in Table 4. The percentage split between training/testing samples varied between 60/40%, 70/30% and 80/20%. The results in all cases were comparable, so for convenience only the 60/40% case is provided here. From the results it is evident that the 11 chosen time-domain features, also known as classical approach features [20], show very high performance. As a contrast, this observation cannot be reaffirmed for the 1st derivatives of all attributes for the speaker independent cross utterance case. With this in mind it can be established that the changes of the features in this scenario are not as emotionally dependent as the features themselves.

k-NN classifier using 11 time-domain features – 60/40 split				
act \ det	H	A	F	N
H	47	3	0	0
A	2	47	0	0
F	0	0	45	0
N	0	0	0	48

Correctly Classified Instances	187	97.40 %
Incorrectly Classified Instances	5	2.60 %
Kappa statistic	0.9653	-
Mean absolute error	0.0180	-
Root mean squared error	0.1135	-
Relative absolute error	4.79 %	-
Root relative squared error	26.2 %	-
Coverage of cases (0.95 level)	97.4 %	-
Mean rel. region size (0.95 level)	25 %	-
Total Number of Instances	192	-

Table 4: Confusion matrix for HAFN using all 11 time-domain features.

DISCUSSION OF RESULTS

Figure 1 displays the mean of each attribute value vs. its rate of change for the speaker independent cross utterance case. The results show that the features used in this study are distinctly separable leading to the 97% class recognition rate. Their derivatives, on a cross utterance level however, were not as prominent since classification on the rate of change was not as separable as the attribute vectors were. This notion is valid across all emotions subject to this study (HAFN). In support, it is observed from the plots in Figure 1 that the average parametric vector for each emotion is clearly different. It is observed that F_p^{F} exhibits flatter pattern than F_p^{HAN}. While F_p^{H} looked somewhat similar to F_p^{A} the two emotion domains A and H were different at F_p^{HA}*[meanE, stdE, maxE]* and their durational attributes F_d were as well. All pitch values were considerably higher in F_p^{HA} as compared to any other class F_p^{FN} except *[stdP]*. Moreover the top four pitch elements F_p^{F}*[meanP, medP, stdP, maxP]* in emotion domain F had a very flat shape, more so than in all the rest of the emotion domains. Continuing our observations further we noted that F_p^{N}*[maxfrP]* as well as F_e^{N}*[meanE, stdE]* for emotion domain N had very distinctive average values in contrast to F_p^{HAF}*[maxfrP]* and F_e^{HAF}*[meanE, stdE]* respectively. Additionally, emotion class F had unique mean values F_e^{F}*[maxE]*, which made it more characteristic than the other emotion domains. The classification displayed in Table 4 undoubtedly shows that using simple features from the time-domain is beneficial in a speaker independent cross utterance situation.

Finally, the applied techniques in this work can improve the search and validation of appropriate content based on automated approach connected to human emotion in speech. This in turn will assist any given digital asset ecosystem that has vast amounts of multimedia content to choose from. Different types of data formats existing in a digital cultural ecosystems, as well as metadata description of cultural resources, can be analyzed using this method for speech and emotion recognition. It is visible from the results that the chosen four emotions are highly separable in the feature domains of choice used in this study, which can lead to more effective solution in developing smart new methods when applied to multifunctional digital content ecosystem. We plan to apply our method for a particular Smart multifunctional digital content ecosystem while pursuing the goal of creating a practical user-friendly application. This in turn will assist the search and

discovery of original Bulgarian historical art when used as an interface and applied to vast media databases containing art, paintings and icons.

Acknowledgements. This work is partly funded by Bulgarian NSF under the research project No. DN02/06/15.12.2016 “Concepts and Models for Innovation Ecosystems of Digital Cultural Assets", Competition for financial support of fundamental research – 2016.

REFERENCES

[1] Eckman P., 1992. An Argument for Basic Emotions. Cognition and Emotion, Vol. 6 (3/4), pp. 169-200.

[2] Cowie R. and Cornelius R., 2003. Describing the Emotional States that are Expssed in Speech. Speech Communication, Vol. 40, pp. 5-32.

[3] Noda T., Yano Y., Doki S., and Okuma S., 2006. Adaptive Emotion Recognition in Speech by Feature Selection Based on KL-divergence. IEEE International Conference on Systems, Man, and Cybernetics in Taipei, Taiwan, October 8-11 2006, pp. 1921-1926.

[4] Ververidis D. and Kotropoulos C., 2006. Emotional Speech Recognition: Resources, Features, and Methods. Speech Communication, Vol. 48, pp. 1162-1181.

[5] Iliev A., Zhang Y., and Scordilis M., 2007. Spoken Emotion Classification Using ToBI Features and GMM, IEEE 6th EURASIP Conference Focused on Speech and Image Processing, pp. 495-498.

[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., Monograph: “Emotion Recognition From Speech”, Lambert Academic Publishing, 2012.

[8] Iliev, A.I., “Emotion Recognition in Speech using Inter-Sentence Time-Domain Statistics”, IJIRSET International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 3, pp. 3245-3254, March 2016.

[9] Witten, I., Frank, E., “Data Mining: Practical machine learning tools and techniques”, Second edition, Morgan Kaufmann, 2005.

[10] Cornelius R., 1996. The Science of Emotion. Research and Tradition in the Psychology of Emotion. Upper Saddle River, NJ: Prentice-Hall, pp. 260.

[11] Cowie R. and Cornelius R., 2003. Describing the Emotional States that are Expressed in Speech. Speech Communication, Vol. 40, pp. 5-32.

[12] Bhatti M., Wang Y., and Guan L., 2004. A Neural Network Approach for Human Emotion Recognition in Speech. IEEE International Symposium on Circuits and Systems, Vancouver BC, pp. 181-184.

ABOUT THE AUTHORS

Assoc. Prof. Alexander I. Iliev, PhD, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences; Assistant Prof. Computing and New Media Technologies, University of Wisconsin, Stevens Point, WI, USA; Lecturer Data Science and Business Intelligence, University of California Berkeley Extension. E-mail: al.iliev@math.bas.bg

Prof. Peter Stanchev, PhD, Kettering University, Flint, Michigan, USA; Professor and Chair of Department at the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria. E-mail: pstanche@kettering.edu