International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

Wide Range Features-Based On Speech Emotion Recognition for Sensible Effective Services

V. Ramesh

Assistant Professor, CSE Department, Sri Indu College of Engineering and Technology, Hyderabad, Telangana, India

ABSTRACT

Speech emotion recognition from speech signals is a noteworthy analysis with many applications like sensible healthcare, autonomous voice response systems, assessing situational seriousness by caller emotive state analysis in emergency centers, and alternative sensible emotive services. During this paper, we have a tendency to present a study of speech emotion recognition supported the options extracted from spectrograms employing a wide range convolution neural network (CNN) with rectangular kernels. Typically, CNN have square shaped kernels and pooling operators at varied layers that are suited to second image information. However, just in case of spectrograms, the data is encoded in a very slightly very different manner. Time is diagrammatical on the x-axis and y-axis shows frequency of the speech signal, whereas, the amplitude is indicated by the intensity value within the spectrograph at a selected position. To research speech through spectrograms, we propose rectangular kernels of variable shapes and sizes, at the side of max pooling in rectangular neighborhoods, to extract discriminative options. The projected theme effectively learns discrimination options from speech spectrograms and performs higher than several state-of the-art techniques once evaluated its performance on emo-db and sample speech data set.

Keywords: Speech Emotion Recognition. Convolution Neural Network. Spectrogram, Rectangular Kernels.

I. INTRODUCTION

With the event of computing, artificial intelligence and human-computer interaction technology, in search of more friendly and vivid human-computer interaction, which needs a pc with the thinking and perception ability that's just like human, in order that the pc will have a lot of humanized functions, and one in every of the most necessary steps is that we want to form computers perceive human emotions. Speech, as a main approach of human communication, will transmit a range of data. In addition, currently the event of the voice device is extremely mature in order that it is a lot of convenient to amass speech signal, and these factors have created the speech feeling recognition become necessary in human-computer interaction.

The purpose of speech feeling recognition is to form PC discover the people's current feeling statement from an individual's voice signal and perceive people's emotional thinking, in order that the communicate between the pc and people in general may be a lot of friendly and humane. Specifically, we will extract the person's spirit options from the speech signal by computers, and analyse and study the human feeling states and their changes, then decide the link between human voice expression and their own emotions, which might facilitate evaluate.

Overall, the method of speech emotion recognition is as follows: speech emotional knowledge acquisition, feeling feature extraction and emotion recognition, the diagram of the system is shown in Figure 1:

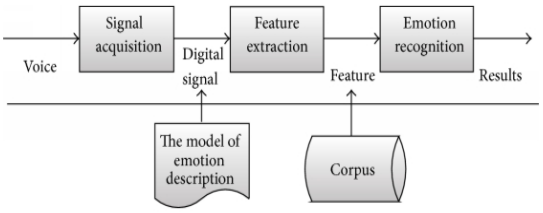

Figure 1. Block diagram of speech emotion recognition

Speech signal is that the most natural, intuitive, and quickest means that of interaction among humans. However, exploitation speech signals to interact naturally with machines need lots of efforts. Significant progress has been created within the recent years in speech recognition and speaker recognition. Speech carries rather more info than spoken words and speaker info. Speech emotions recognition (SER) has been an active area of analysis wherever it aims to alter speech analysis systems to acknowledge the effective state of the speaker [41]. during this context, researchers are effort to form machines understand our emotional states. To extract spirit of a speaker from his/her speech, SER exploitation discrimination options could be a viable answer. However, finding effective and noticeable options for SER could be a difficult task [14]. within the recent past, feeling detection from speech signals has gained abundant attention and is a vital analysis space to form human machine interaction a lot of natural.

With the advancements in technology, there is growing interest in developing affection interaction modes, giving rise to sensible affective services [51]. Hence, applications of SER are increasing at a fast pace. Bjorn Schuller et al. [43] have advised usage of SER inside automotive environments for AN in-car board system wherever ways may be initiated by determinant the status of the motive force to the system. SER will function a necessary element in developing sensible affection services for healthcare, surveillance, human-machine-interaction, audio forensics [27], and affection computing. for example, it may be used as a diagnostic tool for healer [17] and to assess situational seriousness in emergency centers by analyzing the affection state of the callers through their speech [2, 6, and 50]. It can even facilitate in automatic translation systems wherever distinguishing emotion of the speaker plays a significant role in communication between parties. It will collect the supporting analysis information from the emotions of pilgrims from their traditional speaking inside pilgrim's journey services serving to in up information visual image touching serious higher cognitive process [20]. SER will act as a very important tool in serving to mechanically perceive people's physical interactions in numerous dense crowds that is troublesome to try to with manual strategies [25, 26, and 28]. any SER will facilitate in assessing voices of individuals in dense crowds for violent and aggressive behaviors prediction in police work streams [1, 10, and 36]. It can even be employed in supporting information verification theme for input from auto sensor device for proper information gathering [5].

SER has been a full of life space of analysis wherever a spread of approaches are given. vital work has been disbursed in affection options extraction from speech signals and economical classification. Major issue featured by SER systems is that the detection of affect-oriented discriminative options to represent emotional speech signals. Recently, deep neural networks (DNNs) are tested to extract high-level options from raw speech signals that exhibited interesting and extremely acceptable results. Effectively modeling speech signals for classifiers to expeditiously classify emotions may be a vital style issue for SER [14]. During this paper, we tend to gift a convolution neural spec with rectangular kernels and changed pooling strategy to find emotional speech. Furthermore, we tend to conjointly measure the projected technique to acknowledge emotions in emergency calls that are infested with background and poor voice quality, creating SER even tougher. Through intensive experiments, we tend to show that the proposed theme significantly outperforms existing overhand options primarily based SER approaches on the two difficult datasets.

II. SPEECH EMOTION FEATURES

In the method of speech emotion recognition, speech feature is incredibly necessary. It ought to be ready to replicate the speaker's emotion all right and avoid the consequences from alternative components, like the voice of various speakers and therefore the content of speech. It has been found that phonetic options are often divided into 3 categories: prosodic features, the spectrum characteristics and other characteristics.

A. Prosodic Features

Prosodic features can better transfer emotion information in speech, so it has been widely used in speech emotion recognition, and the classical prosodic features are mainly as following: 1) temporal features, such as the longest time to talk; 2) Pitch frequency features, such as the linear prediction coefficient of the

pitch frequency; 3) energy; 4) formant, such as first-order, second-order formant and formant bandwidth; 5) glottal parameters; 6) the phrase, the phoneme, the word, etc.; 7) time structure.

B. Spectrum Features

Feature is usually a short time representation of speech signal. There are some traditional spectral features, such as short time Fourier transform, linear prediction coefficient (LPC), Mel frequency cepstral coefficient (MFCC), line spectrum pair (LSP), perceptual linear prediction cepstral coefficients (PLP), the One-Sided Autocorrelation linear prediction coefficient (OSALPC), short coherence (SMC), linear prediction cepstral coefficients (LPCC) and the One-Sided Autocorrelation linear prediction cepstral coefficient (OSALPCC), zero crossing amplitude peak (ZCPA). Researchers have also proposed some new features, such as performing wavelet transform on each frame of speech signal, then calculating the Fourier transform on its result.

C. Other Features

Prosodic features and spectral features are the most popular features. In addition, there are several common characteristics in speech emotion recognition. These characteristics are: 1) speech feature based on Teager energy operator (TEO). It's able to extract the nonlinear features from the speech signal; 2) speech features based on empirical mode decomposition (EMD). It can be used to analyze the nonlinear and non-stationary speech signals; 3) speech features based on fractal dimension; 4) features based on the deep learning. Traditional features extracted by imitating human auditory effect are often difficult to apply to complex acoustic environment. In the case of low SNR, the effect will be a sharp decline. The features also contain irrelevant information, which affect the accuracy of its direct application in speech recognition.

III. Related Works

Typical SER consists of two main parts 1) a process unit that extracts the most appropriate options from speech signals and 2) a classifier to acknowledge the hidden emotions in speech victimization its options vectors. This section provides a fast summary of existing feature extraction strategies and classification ways.

Common challenges being faced by SER systems include the choice of the speech options, which permit clear discrimination among distinct emotions. However, acoustic variability thanks to the variation of various speakers, speaking designs, speaking rates, and very different sentences directly have an effect on extracted options like pitch and energy contour [3, 7]. To tackle this issue, one potential approach is to divide speech signals into multiple little chunks called frames and construct a feature vector for every frame. For instance, building delivery feature vector for every frame like pitch and energy [40, 48]. Furthermore, world options are often extracted from the full speech auditory communication, which offers lower dimensionality details as compared to native options extracted from every frame, thereby reducing computations. Moreover, it is potential that a selected auditory communication has quite one feeling; every emotion corresponds to very different frame of the spoken auditory communication. Additionally to the present, detecting boundaries of such frames is tough as a result of expression of bound feeling varies from speaker to speaker, cultural variations, and variations in environmental conditions. In literature, most experiments were conducted in monolingual feeling classification environment, wherever the cultural variations among speakers were unnoticed.

Recently, unsupervised feature learning techniques have shown improved results for automatic speech recognition system [52] and image understanding [8, 19]. Stuhlsatz et al. [45] planned a way that yielded improved ends up in each weighted and un-weighted recall by using generatively pre-trained artificial neural network to construct low dimensional discriminative options vector in a very multi-emotion corpora. solon and Kim [42] used deep belief network for emotional music recognition. Their methodology aimed to learn high-level options from magnitude spectra directly as compared to hand-loomed options. The authors in [12, 15, 16, 19, 21–23, 49] replaced a group of Gaussian mixtures by single context dependent DNN using sort of massive scale speech task. Woolmer et al. [48] planned a way to research back and forth speech utterance-levels and used this analysis to predict feeling for a given utterance.

Different types of classifiers are used for SER including hidden mathematician model (HMM) [30, 39], Gaussian mixtures model [53], support vector

machine (SVM) [33], artificial neural network [18], K-nearest neighbor [37] and lots of others [38]. Among these, SVM and HMM are the foremost widely used learning algorithm for speech connected applications [30, 35, 47, 49]. However, experiments show that every classifier is domain dependent in terms of accuracy and the quality of information. Apart from single classifiers, associate degree mass system of multiple classifiers has conjointly been studied for SER for up accuracy [32].

Accurate and economical SER systems will well improve emotional sensible services. With the fast adaptation of end-to-end procedures for classification tasks victimization deep learning algorithms, it becomes imperative to explore these hierarchical architectures for the task of SER on extremely difficult datasets. The strength of those end-to-end learning strategies belong the automatic extraction of discriminative options for economical classification for a spread of information. Dennis et al. [13] planned a novel feature extraction methodology for classification of sound events. They extracted the visual signature from sound's time-frequency illustration (i.e., spectrograms). They tested their methodology on an info consisting of 60 sound categories. They claimed a stimulating improvement over different strategies in conditions with couple. Deng Li et al. [11] explored a layer-by-layer learning strategy of patches of speech spectrograms for Multimedia Tools Appl.

Training a multi-layer generative model. Qirong Mao et al. [34] introduced feature learning to SER by learning affected-salient options victimization CNN. They used public feeling speech databases with different languages. With relation to speaker variation, language variation and environmental noise, they achieved high results with learned options compared to different established feature representations. There are several strategies to perform feeling recognition victimization CNNs, but few of them are victimization spectrograms to acknowledge emotions from speech that so could be a new approach in SER. Among the strategies victimization spectrograms for SER, a number of them have used an additional classifier at connected layer that will increase the procedure complexness of the general model. for instance, in [34] the effect-salient feature block obtained from the ultimate feature vector is then passed to a SVM categoryifier [31] to get the feeling class of the speech auditory communication. The second reason that makes our work very different from the present works is that we have introduced the employment of rectangular kernels, which permit extraction of meaty options from spectrograms. The oblong kernels and pooling operations area unit used keeping seeable the format of knowledge conferred in spectrograms. Lastly, we have used a changed Alex Net design that uses a comparatively easy layout, compared to fashionable architectures and is a smaller amount susceptible to over fitting with limited training data.

IV. Proposed Method

The proposed framework utilizes feature-learning scheme powered by a discriminative CNN using spectrograms to recognize the spirit of the speaker. The most parts of the planned framework are explained within the subsequent sections.

A. Spectrograms extraction from speech

A spectrogram represents the strength or loudness of a sign over time at completely different frequencies in a very explicit undulation. With the energy strength at a specific region, we are able to additionally see the variation within the energy over time. In general, spectrograms are wont to see the frequencies in continuous signals. It is a graph with two geometric dimensions during which time is shown on the horizontal axis, whereas the vertical axis represents frequency, and therefore the intensity or colour of every purpose within the image corresponds to amplitude of explicit frequency at explicit time.

Short term Fourier remodel (STFT) is typically applied to an electronically recorded sound, to come up with spectrograms from the sign. mistreatment quick Fourier remodel (FFT) for generating the spectrogram may be a digital method. to find frequencies at every purpose within the speech signal, little window is moved over the signal and FFT is computed for the signal inside every window. For a given spectrogram S, the strength of a given frequency part f at a given time t within the speech signal is described by the darkness or color of the corresponding purpose S(t,f).We extracted spectrograms as shown in Fig. one by using STFT for every audio go in the dataset. Figure one contains sample spectrograms for every of the seven emotions in Emo-DB dataset.

In the gift work, we extracted spectrograms from every individual file then we tend to split the spectrogram into multiple smaller spectrograms with an overlap of fifty. This overlap served 2 functions. First, it allowed USA to simulate continuity in the process pipeline. Secondly, it caused a rise within the variety of

spectrograms that permits USA to effectively train or fine-tune a robust deep CNN. The ensuing photograph pictures had dimensions sixteen × 256, that were resized to 256 × 256 for input to the CNN.

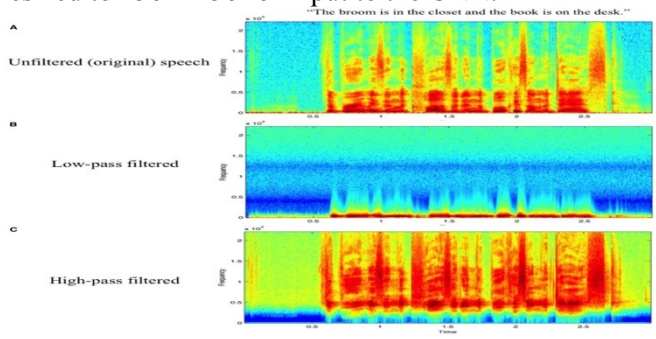

Figure 2. Spectrograms for various emotions

B. Convolutional neural network

Convolutional Neural Networks are very like ordinary Neural Networks from the previous chapter: they're created from neurons that have learnable weights and biases. every vegetative cell receives some inputs, performs a real and optionally follows it with a non-linearity. the entire network still expresses one differentiable score function: from the raw image pixels on one finish to category scores at the opposite. and they still have a loss perform (e.g. SVM/Softmax) on the last (fully-connected) layer and every one the tips/tricks we tend to developed for learning regular Neural Networks still apply.

So what will change? ConvNet architectures create the express assumption that the inputs are pictures that allows United States to write sure properties into the design. These then create the forward perform a lot of economical to implement and immensely reduce the quantity of parameters within the network.

Typically, a CNN may be a hierarchical neural network that consists of a stack of convolution layers followed by pooling layer that performs feature extraction by reworking a picture (i.e. spectrogram) to the next level abstraction in a very layer by layer manner. The initial layers carries with it straightforward options like raw image pixels and edges, the upper layers contain native discriminative options, whereas the last dense (fully connected) layer derives a worldwide representation from the native convolutional options that is then fed to a Softmax categoryifier to come up with probabilities for every class. A Convolutional layer applies convolution filters on l the little portion of the input image and produces single price within the output feature map by activity real and summation operations on these small regions. Every Convolutional kernel generates a feature map wherever the activation values correspond to the presence of specific options. Many feature maps are generated inside every convolution layer. Between serial Convolutional layers, a pooling layer is applied that controls over fitting and reduces computations within the network. the foremost usually used pooling formula is GHB pooling that keeps the utmost price and discards all alternative values in a very native neighbourhood. Connected layers use many wide filters to method many complex parts within the input layer. Each node within the connected layer is connected to each node in preceding layer. the choice of acceptable kernels shapes and sizes, and pooling neighbourhoods is vital to the success of those models.

C. Proposed model

The proposed CNN framework is shown in Fig. 2, that has an input layer, 5 Convolutional layers, 3 pooling layers, 3 absolutely connected layers, and a Softmax layer. The spectrograms generated from emotional speech signals (16 × 256, resized to 256 × 256) are input to the CNN. Convolutional kernels square measure applied to the input within the initial layers to extract feature maps from these spectrograms. the primary Convolutional layer C1 has ninety six kernels of size fifteen × three that square measure applied with stride setting of (3 × 1) pixels.

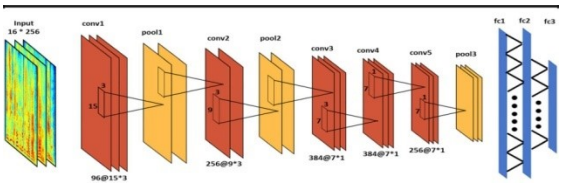

Figure 3. Proposed CNN architecture with rectangular kernels

This network was designed keeping in view the format of knowledge encoded within the spectro-grams. Every input spectrogram corresponds to a shorter sample of the input speech wherever the frequencies and amplitudes area unit encoded. Within the lower layers, the kernels have bigger heights and comparatively lower widths so they will capture native options effectively from the neighborhood. Within the resultant layers, each

the peak and dimension area unit reduced however the form of the kernels still stays rectangular. It helps to construct effective native receptive fields for spectrograms. The salient options of this design area unit the oblong formed kernels, strides, and pooling neighborhoods, that build it potential for the CNN to effectively capture discriminative options from spectrograms.

D. Model coaching

The planned CNN design was enforced in Caffe [24], victimization NVidia DIGITS five.0 as frontend [54] for coaching and confirming models. MATLAB was wont to generate spectrograms from every feeling within the dataset. The spectrograms of size sixteen × 256 were generated with a five hundredth overlap. Around 1500 spectrograms were generated for every feeling within the dataset. Overall, quite ten, 000 spectrograms were generated for all the audio files within the dataset. These spectrograms were divided for coaching and testing in such the way that seventy fifth of knowledge the info the information} was used for coaching and twenty fifth data was wont to validate the performance of the model. A 5-fold cross validation was used for all experiments.

The coaching method was last thirty epochs with a batch size of 128. we have a tendency to set the initial learning rate to zero.01 with a decay of one when each ten epochs. one NVidia GeForce GTX Titan X GPU with twelve GB on-board memory, was wont to train and later fine-tune the models. the most effective accuracy was achieved when twenty nine epochs. A loss of zero.8066 was obtained on the coaching set, whereas on the take a look at set a loss of one.0695 was determined.

A single sound recording sometimes consists of the many spectrograms (as shown in Fig. 3). it's price mentioning that if a lot of that half-hour spectrograms for one file were properly classified, then the prospect of properly predicting the whole file are considerably high. each the result per image and per file area unit shown within the result section intimately.

E. Feeling prediction victimization majority option

Individual spectrograms generated for audio stream area unit input to the trained CNN for prediction. Such a small fragment of speech might not be comfortable to properly predict any emotions. Hence, prediction results from multiple spectrograms area unit combined victimization majority option theme for reliable prediction performance. For the speech stream, an oversized pic is generated that is later segmental into multiple short spectrograms as shown in Fig. 4. Predictions area unit obtained by victimization the trained model for every generated pic. Chances of seven completely different emotions area unit obtained from the Softmax layer of the model for the Emo-DB dataset. Similarly, for the Korean speech dataset, constant design was wont to predict emotional or normal speech. The general prediction scores for these emotions area unit then obtained by employing a majority option theme wherever the foremost frequent label is allotted to the speech stream or a group of streams if the recording is drawn-out and will contain multiple emotions. within the current state of affairs supported the collected proof from multiple spectrograms if roughly 25–30% predictions of individual spectrograms for one audio file area unit created properly, then there exists a good probability that the particular feeling are predicted accurately.

V. CONCLUSION

In this paper, we tend to present technique to recognize emotions in speech using Convolutional neural network with rectangular kernels. Speech signals are represented as spectrograms that area unit generated with a five hundredth overlap. Generated spectrograms were resized to fit the requirements of the CNNs throughout coaching and analysis. 2 different CNNs were trained on the spectrograms having different kernel sizes and pooling approaches. Within the initial CNN, the default design, kind of like AlexNet was used. Whereas, the second CNN was obtained by modifying the kernel sizes and pool neighborhoods from sq. to rectangular, to form it additional appropriate for spectrograms. Each the CNNs were trained using identical dataset and similar parameters. for every pic, the trained CNNs generated possibilities. For classifying a specific speech phase, majority choice theme was used.

VI. REFERENCES

[1]. Abdelgawad H, Shalaby A, Abdulhai B, Gutub AAA (2014) Microscopic modelling of large-scale pedestrian-vehicle conflicts in the city of Madinah, Saudi Arabia. J Adv Transp 48:507-525

[2]. Ahmad J, Muhammad K, Kwon S-I, Baik SW, Rho S (2016) Dempster-Shafer Fusion Based Gender Recognition for Speech Analysis Applications. In: Platform Technology and Service (PlatCon), 2016 International Conference on, pp 1-4

[3]. Ahmad J, Sajjad M, Rho S, Kwon S-I, Lee MY, Baik SW (2016) Determining speaker attributes from stress-affected speech in emergency situations with hybrid SVM-DNN architecture. Multimed Tools Appl 1-25. https://doi.org/10.1007/s11042-016-4041-7

[4]. Ahmad J, Fiaz M, Kwon S-I, Sodanil M, Vo B, Baik SW (2016) Gender Identification using MFCC for Telephone Applications-A Comparative Study. International Journal of Computer Science and Electronics Engineering 3.5 (2015):351-355

[5]. Aly SA, AlGhamdi TA, Salim M, Amin HH, Gutub AA (2014) Information Gathering Schemes For Collaborative Sensor Devices. Procedia Compute Sci 32:1141-1146

[6]. Badshah AM, Ahmad J, Rahim N, Baik SW (2017) Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In: Platform Technology and Service (PlatCon), 2017 International Conference on, pp 1-5

[7]. Banse R, Scherer KR (1996) Acoustic profiles in vocal emotion expression. J Pers Soc Psychol 70:614

[8]. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798-1828

[9]. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of German emotional speech. In: Interspeech, pp 1517-1520

[10]. Curtis S, Zafar B, Gutub A, Manocha D (2013) Right of way. Vis Compute 29:1277-1292

[11]. Deng L, Seltzer ML, Yu D, Acero A, Mohamed A-R, Hinton GE (2010) Binary coding of speech spectrograms using a deep auto-encoder. In: Interspeech, pp 1692-1695

[12]. Deng J, Zhang Z, Marchi E, Schuller B (2013) Sparse autoencoder-based feature transfer learning for speech emotion recognition. In: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, pp 511-516

[13]. Dennis J, Tran HD, Li H (2011) Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Process Lett 18:130-133

[14]. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn 44:572-587

[15]. Engberg IS, Hansen AV, Andersen O, Dalsgaard P (1997) Design, recording and verification of a danish emotional speech database. In: Eurospeech

[16]. Eyben F, Wöllmer M, Schuller B (2009) OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. In: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, pp 1-6

[17]. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes M (2000) Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng 47:829-837

[18]. Gharavian D, Sheikhan M, Nazerieh A, Garoucy S (2012) Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Compute & Applic 21:2115-2126

[19]. Guo Z, Wang ZJ (2013) An unsupervised hierarchical feature learning framework for one-shot image recognition. IEEE Trans Multimedia 15:621-632

[20]. Gutub A, Alharthi N (2011) Improving Hajj and Umrah Services Utilizing Exploratory Data Visualization Techniques. Inf Vis 10:356-371

[21]. Guven E, Bock P (2010) Speech emotion recognition using a backward context. In: Applied Imagery Pattern Recognition Workshop (AIPR), 2010 I.E. 39th, pp 1-5