Speech Emotion Recognition Using Deep Learning Techniques

Apoorva Ganapathy

Senior Developer, Adobe Systems, San Jose, California, **USA**

ABSTRACT

The developments in neural systems and the high demand requirement for exact and close actual Speech Emotion Recognition in human-computer interfaces mark it compulsory to liken existing methods and datasets in speech emotion detection to accomplish practicable clarifications and a securer comprehension of this unrestricted issue. The present investigation assessed deep learning methods for speech emotion detection with accessible datasets, tracked by predictable machine learning methods for SER. Finally, we present-day a multi-aspect assessment between concrete neural network methods in SER. The objective of this investigation is to deliver a review of the area of distinct SER.

Keywords: Deep learning, LSTM, emotional speech database, speech emotion recognition

12/30/2016 **Source of Support:** Nil, **No Conflict of Interest:** Declared

This article is is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.
Attribution-NonCommercial (CC BY-NC) license lets others remix, tweak, and build upon work non-commercially, and although the new works must also acknowledge & be non-commercial.

INTRODUCTION

The mechanism set up to identify emotions and recognise speech signals is known as speech emotion recognition. This is very vital progress in human to computerized communication: Human-machine communication consists of about five basic areas of work. This include; a study into interactional software and hardware, study bothering on model matching, studies involving task at level, studies regarding design, and those consisting of impact to the organization (Booth, 1989). Harper et al. (2008) and Cambria et al. (2010) opine that the comprehension of one's emotion during communication is essential; first for to understand the discussion at hand and secondly to respond effectively. At the moment, this aspect of computer-human communication is yet to be solved except for a few applications, there is no way out to this challenge. As with all other aspects and emerging issues in machine learning, Speech Emotion Recognition (SER) has begun to surge greater advantage from the application and packages brought by deep learning. Gaussian mixture models (GMM), Support Vector machines (SVM), and Hidden Markov models (HMM) with many precise and preprocessing engineering feature are some of the techniques which Speech Emotion Recognition depended on prior to the massive application of deep learning (Lin and Wei, 2005; Patil et al., 2010; Hassan and Damper, 2010).

Nevertheless, with the emergence and resurgence of deep learning in the public domain and numerous literature, the results relating to the control of the environment now have about a 70% accuracy rate. Automatic Speech Emotion Recognition assists virtual assistants and smart speakers to comprehend their users properly, particularly as they if comprehend bijous words. For instance, the word "really" can form part of a question or applied interactively to understand a fact and emphasize a statement in either negative or positive ways. Nicholson et al. (1999) noted that this tool can also assist in changing from language to language. It can translate languages and possess various methods it can use to explain and reveal the intention of the speakers through speech.

The application of Speech Emotion Recognition is also essential in online discussions and tutorial classes. Comprehension of the emotional status of the student can assist the tool to decide the technique or methodology to be adopted to present the remaining parts of the lesson contents (Schüller et al., 2004). More so, vehicles benefit greatly from Speech Emotion recognition. It is important towards ensuring vehicle safety is well documented. How does it handle vehicle safety? (Schüller et al., 2004). It understands the state of mind or the emotions of the driver and guides against crashes and other road disasters. In the area of therapy, the application of Speech Emotion Recognition helps the therapist to comprehend patients effectively. Their mindset and the underlying emotions are rather hidden (France et al., 2000; Hansen and Cairns, 1995). In stressful and noisy locations such as the cockpits in aircraft, the use of Speech Emotion Recognition assists to identify the emotions of the pilot and its state of mind. Its use is also witnessed in the service industry and e-commerce. With the use of speech emotion recognition tools, call centers can now relate better to customers in the area of alerting and sending prior notice of the caller's state of mind to supervisors and customer service (Petrushin, 2000). Further, the tool of speech emotion recognition has been integrated into movies which are rather interactive in nature. With it, the emotions of the viewers are properly gauged. This can assist the producers of movies to channel, or review such movies to end differently or maintain their ending as planned (Nakatsu et al., 2000).

In other to effectively classify emotions with machine learning algorithms there is a need to acquire a training dataset. For the task of Speech Emotion Recognition, there are three (3) forms of training datasets, simulated, natural, and semi-natural. The data in natural forms are retrieved from audios, and videos either transmitted online, TV, or online. Datasets are also retrieved from call points and similar centers. The Semi-natural data forms are derived by outlining situations for experts' voice actors and requesting them to play them. Simulated datasets which is the third form of the dataset are also the most widely used type and functions almost like the semi-natural. The variation is in the same sentences which are being acted by the voice actors. Primarily, Speech Emotions Recognition adopted the procedures of automatic speech recognition (ASR), and technique based on Support vector machines (SVMs), GMMs, and HMMs were widely used (Hassan and Damper, 2010; Lin and Wei, 2005; Amir et al., 2001; Erden and Arslan, 2011; New et al., 2003; Chavhan et al., 2010).

PROBLEM STATEMENT

These methods required enormous engineering features and any variation in the features would need re-modeling the overall architecture of the technique. Nevertheless, recent development in deep learning applications and methods for Search Emotion Recognition can be varied also. There are numerous literature and studies on the application of these algorithms to understand emotions and state of minds from human speech (Shaw et al.,

2016; Stuhlsatz et al., 2011; Han et al., 2014; Amer et al., 2014). Additionally, to deep learning, neural networks, and application of improvements of long short-term memory (LSTM) networks, generative adversarial models, and lots more, a wave in researches on speech emotion recognition and its application now emerges. It is essential to understand its application and its role towards emotion (Wöllmer et al., 2013; Trigeorgis et al., 2016).

For this reason, the objective of the current paper is to understand deep learning techniques for speech emotion recognition, from databases to models.

This article is structured into 5 sections: Section 1 is the introduction, problem statement, and objectives of the study. In Section 2, the literature review; definition of Speech Emotion Recognition, and the presentation of basic and related researches in the field. In Section 3, the methodology to concluding is scheduled. In the chapter, the result is presented. Section 4 schedules the results and discussion, while the last session takes care of the conclusion and recommendation of the paper and schedules future actions in the field.

LITERATURE REVIEW

For easy understanding of the context of the subject matter, it is important to balance other signals and what we listen to from the converser. Comprehending the passions of our collocutor whereas interactive is one of the signals. Comprehending the passions concerning the message understanding would be a crucial key to a meaningful dialog (Williams and Stevens, 1972). Together with numerous benefits that could be again by humans on the account of comprehending passions, it can be seen in human-computer communication. A lot could be gained as well with a proper understanding of passions. In the contemporary age, a lot of research has been carried out, impacts, and even competitions aimed at designing manners and approaches to produce such comprehension for computers (Balti and Elmaghraby, 2013; Ujwala et al., 2012; Kaushik et al., 2013; Vadlamudi, 2016; Balti and Elmaghraby, 2014).

Computer algorithms are needed for the easy classification of emotions, and there must be a mathematical model relating them. The standard method demarcated by psychologists according to three (3) processes that produce a three-D galaxy that defines all the passions or emotions. These dimensions are arousal, pleasure, and dominance (Grimm et al., 2008; Truong et al., 2017). A mishmash of these potentials will produce a path that will be in one of the measured passion territories, and according to that, we can document the utmost pertinent passion (Albert, 1996).

Utilizing arousal, pleasure, and dominance, any kind of emotion can be described, nevertheless, that kind of deterministic approach can be somehow difficult to operate in machine learning. However, when dealing with machine learning (ML), classically, we utilizing cluster samples and statistical models into any of the so-called qualitative feelings such as happiness, anger, sadness among others. For one to categorize and group easily any of the talk about feelings, one has to pattern those employing characteristics isolated from the speech; this is achieved by means by digging out diverse classes of voice quality, prosody, and spectral structures (Vadlamudi, 2015).

There are series of benefits associated with these classifications in identifying weaknesses and some emotions in classifying others. The prosody properties utilizing aimed at essential frequency, duration, speaking rate, and intensity, could not be used to confidently distinguish happy and angry reactions from one another (Movva et al., 2012).

However, voice quality properties are typically prevailing in the identification of feelings of the same talker. Although, it varies from talker to talker, and this is becoming difficult in employing them in a talker self-determining setting (Gobi and Chasaide, 2003). Spectral structures have been widely studied to develop passions from speech. The instantaneous benefit that can be likened to prosody potentials is that it can assertively differentiate happiness from anger. Thus, the main point of contention is that shift of the formant and magnitude for similar feelings differ through diverse vowels, and this contributed to the complexity of a feelings detection system, and it requires to be speech pleased cognizant (Vlasentho et al., 2011).

About any of those structure classifications, as stated before, there are many typically structure demonstrations. Prosody structures are commonly presented by essential frequency and process link to the rate of speaking (Lee and Narayanan, 2005), and spectral structures are typically termed by employing one of the cepstrum-oriented illustrations existing. Generally, linear prediction cepstral coefficients (LPCC) or Mel-frequency cepstral coefficients (MFCC) are employed, and some papers, formants, spectral structures, and any other information are utilized altogether (Schüller and Rigoll, 2006). Voice quality structures are commonly explained by normalized amplitude quotient (NAQ), jitter, and shimmer (Lugger and Yang, 2007).

Two main methods in speech emotion recognition are so far classified, whichever detecting according to the 3D of emotion or detection according to statistical model recognition methods for the qualitative so-called emotions. The first method deals with the calculation degree of relationship between the given passion and signal, dominance, and arousal, and before utilizing a hierarchical classifier, the composite emotion is recognized. The second approach is achieved utilizing statistical model detection techniques like Gaussian mixture model (Erden and Arslan, 2011), hidden Markov model (Nwe et al., 2003), deep neural network (Amer et al., 2014), artificial neural network (Amir et al., 2001), support vector machine (Hassan and Damper, 2010), and genetic algorithm (Philippou-Hübner et al., 2010). Owing to the peculiarity of speech emotion recognition (SER) in human-computer interface and the advancement of artificial intelligence (AI) structures, there are thousands of surveys and publications on speech emotion recognition. Thus, this subsection will consider the most topical related literature to SER.

Anjali et al. (2020) review speech emotion recognition approaches. This review cover 2009 to 2018 and several features applied in SER. Despite its limitations, it can still be considered as a starting point. Also, Paruchuri (2015) documented a review on the significance of speech emotion features such as noise reduction and dataset. The importance of diverse classification methods including support vector machine and hidden Markov model. Identification of various features associated with SER was considered the strength of the study while leaks of modern approaches were identified as a limitation. Also, they suggested recurrent and convolutional neural networks as a deep learning technique.

METHODS

Based on the importance of speech emotion recognition (SER) in the human-computer interface and for us to achieve the objective of this study, this section will be divided into two parts; the speech emotion datasets categories, their features, and examples, and SER techniques.

Speech emotion datasets

For every ML assignment, there is a need to have a working outset of samples, speech emotion recognition is not poles apart from the rest. The procedure of producing working out datasets for speech emotion recognition requires human means to tag the samples manually, and diverse individuals recognize emotions otherwise.

For instance, one may label an emotional voice as being angry and another may recognize it as excitement. This obscurity agents to tag the samples, there is needs to have more than an agent labeling each voice and then setting a system to select the tags for existing samples assertively.

There are 3 types of datasets mainly considered for SER; natural, semi-natural, and simulated speech pools. The natural databases are collected from YouTube, TV shows, Call centers, and the likes and then tagged the emotions by human audiophiles whereas semi-natural is generated from actors or individuals assigned to read a scenario comprising diverse emotions. The simulated databases are produced by competent speakers reading the manuscript with dissimilar emotions (Douglas-Cowie et al., 2000). These datasets and their features are summarized in Table 1 summarizes examples for each type of dataset.

Table 1: The speech emotion datasets categories, their features, and examples

	Natural	Semi-Natural	Simulated
Natural emotions	+	+	-
Comprises traditional information	+	+	-
Comprises conditional information	+	+	-
Separable and discrete emotion	-	-	+
Distinct emotions at a moment	-	+	+
Widely utilized	-	-	+
Uniform	-	-	+
Stress-free to model	-	-	+
Inter corpora outcomes are liken	-	-	+
An outsized number of emotions	-	+	+
Utilized in real-world emotional structures modeling	-	-	+
Controlled copyright and privacy	-	+	+

Key: (-) denote the features that are absent; (+): denote the features are present

Deep Learning Methods for Speech Emotion Recognition Improvement

One main restriction in emerging an influential speech emotion detection method accomplished of managing everyday situations is the strangeness of the sequence and assessment datasets and deficiency of broad view. To astound these glitches and recover the competence of the speech emotion recognition procedures, there are numerous approaches combined as trappings to the base wedges employed in speech emotion recognition. Some of the methods of deep learning employ in SER improvement include generative adversarial networks (GANs), multitask learning, autoencoders, attention mechanism, and transfer learning. Their role cannot be overemphasized.

RESULTS AND DISCUSSION

Table 4 summarizes a brief assessment of some of the approaches mentioned in this article providing the paper title, year of publication, the methods employed, the features and datasets employed and percentage accuracy obtained for the individual datasets.

Table 4: Comparison of some Algorithms mentioned in this study

Work Title	Method and number of strata	Features	Databases and Precision
LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework (Wöllmer et al., 2013)	LSTM/1 BLSTM/1	• Loudness, ZCR, Energy in 1–4 kHz, 250–650 Hz • 25%, to 90% spectral roll-off points, entropy, flux, skewness, variance, • Psychoacoustic intelligence, harmonicas, 10 MFCCs • F0 (SHS followed by Viterbi smoothing) • Voicing, shimmer (local), jitter, delta jitter • Logarithmic Harmonics-to-Noise • Ratio (logHNR)	SEMAINE: 71.0%
Adieu, Features? End-To-End Speech Emotion Recognition Using A Deep Convolutional Recurrent Network (Trigeorgis et al., 2016).	LSTM/4; DCNN	PCM	RECOLA: 68.4%
On Enhancing Speech Emotion Recognition Using Generative Adversarial Networks, (Sahu et al., 2018)	GAN, SVM	1582-dimensional open smile feature space	IEMOCAP:61.29%
Data Augmentation Using GANs for Speech Emotion Recognition, (Chatziagapi et al., 2019)	DCNN (VGG19), GAN/19	128 MFCCs	IEMOCAP: 52.6% Feel-25k: 53.6%
human–computer Using Transfer Learning, (Song et al., 2014)	PCA, TSL LPP,	12 MFCCs and Delta 8 LSF, Intense, Loudness, ZCR •Voice probability, F0, F0 envelopes	EMO-DB: 58.8%
Transfer Linear Subspace Learning for Cross-Corpus Speech Emotion Recognition, (Song 2019)	LDA, TLSL, TSL,	1582-dimensional openSMILE feature space	EMO-DB, eNTERFACE, Aibo: 53.61%
Automatic Speech Emotion Recognition using recurrent neural networks local attention (Mirsamadi et al., 2017)	LSTM, ATTN/4, 3, 3, 4, 4, 4	257-dimensional magnitude FFT vectors F0, voice probability, frame energy, ZCR 12 MFCCs and Delta	IEMOCAP: 62.5%
Improved end-to-end Speech Emotion Recognition using self-attention mechanism and multitask learning (Li et al., 2019)	CNN BLSTM, ATTN/2, 2	800 point STFT Mel scale spectrogram, Deltas, Delta deltas	IEMOCAP: 82.8%
Speech Emotion Recognition based on HMM and SVM, (Lin and Wei, 2005)	HMM	1st and 2nd derivative of F0, 1stderivative of F1 2nd derivative of MBE4, 2nd derivative of MBE5 MFCC	DES: 98.5%
Speech Emotion Recognition Using Support Vector Machines (Chavhan et al., 2010)	SVM	MFCC MEDC	EMO-DB: 93.70%

Emotion Recognition and Classification in Speech using Artificial Neural Networks, (Shaw et al., 2016)	ANN/1	Energy, Pitch, Formants 0 to 4 20 MFCCs	86.86%
Emotion recognition from Marathi speech database using adaptive artificial neural network, (Darekar, and Dhande, 2018)	ANN/1 PSO-FF	MFCC NMF Pitch	RAVDESS: 87.7%
Emotion Detection in Speech Using Deep Networks, (Amer et al., 2014)	CRF/3 CRBM	Spectrum	AVEC: 68.2% VAM: 65.3% SPD: 74.6%

In likening the recommended techniques, interest was on measures rather than unweighted or weighted precision. Thus, very few articles documented other means like F1 grade. Some of the articles mentioned multiple precisions documented for many conditions that it was only the best precision was summarized for individual datasets employed. Close examining of Table 4 shows that among various features employed for speech emotion detection assignment, MFCC was the dominant feature used. Also, open smile features and feeding raw audio data are employed in tropical reports. In previous studies, the majority of the techniques depends on traditional machine learning like SVM and signal handling techniques. Most tropical studies are aiming at deep learning and neural network enhancement that directly associated with the improvement and progress of software and hardware that permits investigators to use and tune refined networks like GAN, LSTM, and VAE. The precisions reported in Table 4 include results with precision far above ninety percent. From the Table, all techniques are employing older datasets like DES and EMO-DB in which both of them have a negligible number of cases.

According to Table 4, there is no ostensible correlation between the composite feature fixed and the precision documented and the suggested techniques possess an essential role in the outcomes. Neogy & Paruchuri (2014) incorporated similar datasets employing EMO-DB, PCM as the feature set for samples of the wave file, and the precision was 97.1%. However, Song et al (2014) with a composite feature fixed, reported precision of 59.7%.

Once more reason could be that EMO-DB has a unit of magnitude lesser number of samples when compared to IEMOCAP, and employing it with deep learning techniques allow it further prone to overfitting. As an outcome of used feature separation and cataloging techniques that are employing signal processing, apart from in just a minimal situation, show general lower precisions. Conversely, the average precision has been improved in recent times. Although, the noise sensitivity and overfitting of deep learning techniques is yet to be addressed and studies are still ongoing to address these problems too.

CONCLUSION

This study has established the two most widely used simulated databases that are EMO-DB and DES. Also, 3 new, freely existing simulated databases in English. In addition to the simulated datasets, IEMOCAP was added, and regularly quoted semi-natural dataset and VAM, a typical natural dataset. This article also discussed various speech emotion recognition available literature. All the most used deep learning techniques for speech emotion detection, starting from DNNs to LSTMs. The most reported limitation is the precision of the detection as well as their performance measure but statistically, self-precision is not an understandable measure of the performance of the system. The increase

of new studies on CNNs demonstrates that they are able of solving the issues of SER by having greater low level and short-term differentiation abilities. The combination of LSTM connections and institution of DC-LSTM systems has facilitated to carry the clarification to a new level and to give the connection long-duration memory to aid detect long term paralinguistic models. Further investigation possibly will consider more vigorous and dataset autonomous answers to enable patterns moving closer to the invention in real life.

REFERENCES

Albert, M. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. Curr. Psychol., 14, 261–292.

Amer, M.; Siddiquie, B.; Richey, C. and Divakaran, A. (2014). Emotion Detection in Speech Using Deep Networks. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014.

Amir, N.; Kerret, O. and Karlinski, D. (2001). Classifying emotions in speech: A comparison of methods. In Proceedings of the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark, 3–7 September 2001.

Balti, H. and Elmaghraby, A.S. (2013). Speech emotion detection using time dependent self-organizing maps. In Proceedings of the IEEE International Symposium on Signal Processing and Information Technology, Athens, Greece, 12–15 December 2013.

Balti, H. and Elmaghraby, A.S. (2014). Emotion analysis from speech using temporal contextual trajectories. In Proceedings of the IEEE Symposium on Computers and Communications (ISCC), Funchal, Portugal, 23–26 June 2014.

Booth, P.A. (1989). An Introduction to Human-Computer Interaction; Psychology Press: Hove, UK.

Cambria, E.; Hussain, A.; Havasi, C. and Eckl, C. (2010). Sentic computing: Exploitation of common sense for the development of emotion sensitive systems. In Development of Multimodal Interfaces: Active Listening and Synchrony; Springer: Berlin/Heidelberg, Germany, pp. 148–156.

Chavhan, Y.; Dhore, M. and Pallavi, Y. (2010). Speech Emotion Recognition Using Support Vector Machines. Int. J. Comput. Appl., 1, 86–91.

Erden, M. and Arslan, L.M. (2011). Automatic detection of anger in human-human call center dialogs. In Proceedings of the Twelfth Annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011.

France, D.J.; Shiavi, R.G.; Silverman, S.; Silverman, M. and Wilkes, D.M. (2000). Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk. IEEE Trans. Biomed. Eng., 47, 829–837.

Gobl, C. and Chasaide, A.N. (2003). The role of voice quality in communicating emotion, mood and attitude. Speech Commun., 40, 189–212.

Grimm, M.; Kroschel, K. and Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In Proceedings of the IEEE International Conference on Multimedia and Expo, Hannover, Germany, 23–26 June 2008.

Han, K.; Yu, D. and Tashev, I. (2014). Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.

Hansen, J.H. and Cairns, D.A. (1995). ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments. Speech Commun., 16, 391–422.

Harper, E.R.; Rodden, T.; Rogers, Y. and Sellen, A. (2008). Being Human: Human-Computer Interaction in the Year 2020; Microsoft Research: Redmond, WA, USA; ISBN 0955476119.

Hassan, A. and Damper, R.I. (2010). Multi-class and hierarchical SVMs for emotion recognition. In Proceedings of the INTERSPEECH 2010, Makuhari, Japan, 26–30 September 2010; pp. 2354–2357.

Kaushik, L.; Sangwan, A. and Hansen, J.H.L. (2013). Sentiment extraction from natural audio streams. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013.

Lee, C.M. and Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. 13, 293–303.

Lin, Y.L. and Wei, G. (2005). Speech emotion recognition based on HMM and SVM. In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; Volume 8, pp. 4898–4901.

Lugger, M. and Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. ICASSP, 4, 17–20.

Movva, L., Kurra, C., Koteswara Rao, G., Battula, R. B., Sridhar, M., & Harish, P. (2012). Underwater Acoustic Sensor Networks: A Survey on MAC and Routing Protocols. *International Journal of Computer Technology and Applications, 3*(3).

Nakatsu, R.; Nicholson, J. and Tosa, N. (2000). Emotion recognition and its application to computer agents with spontaneous interactive capabilities. Knowl.-Based Syst., 13, 497–504.

Neogy, T. K., & Paruchuri, H. (2014). Machine Learning as a New Search Engine Interface: An Overview. *Engineering International*, *2*(2), 103-112. https://doi.org/10.18034/ei.v2i2.539

Nicholson, J.; Takahashi, K. and Nakatsu, R. (1999). Emotion Recognition in Speech Using Neural Networks. In Proceedings of the 6th International Conference on Neural Information Processing (ICONIP '99), Perth, Australia, 16–20 November 1999.

Nwe, T.L.; Foo, S.W. and De Silva, L.C. (2003). Speech emotion recognition using hidden Markov models. Speech Commun., 41, 603–623.

Paruchuri, H. (2015). Application of Artificial Neural Network to ANPR: An Overview. *ABC Journal of Advanced Research*, *4*(2), 143-152. https://doi.org/10.18034/abcjar.v4i2.549

Patil, K.J.; Zope, P.H. and Suralkar, S.R. (2012). Emotion Detection From Speech Using Mfcc and Gmm. Int. J. Eng. Res. Technol. (IJERT), 1, 9.

Petrushin, V. (2000). Emotion in Speech: Recognition and Application to Call Centers. Artif. Neural Netw. Eng. 2000, 710, 22.

Philippou-Hübner, D.; Vlasenko, B.; Grosser, T. and Wendemuth, A. (2010). Determining optimal features for emotion recognition from speech by applying an evolutionary algorithm. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 2358–2361.

Schüller, B. and Rigoll, G. (2006). Timing levels in segment-based speech emotion recognition. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006; pp. 17–21.

Schüller, B.; Rigoll, G. and Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004.

Sepp Hochreiter, J.S. (1997). Long Short-Term Memory. Neural Comput., 9, 1735–1780.

Shaw, A.; Vardhan, R.K. and Saxena, S. (2016). Emotion Recognition and Classification in Speech using Artificial Neural Networks. Int. J. Comput. Appl., 145, 5–9.

Song, P.; Jin, Y.; Zhao, L. and Xin, M. (2014). Speech Emotion Recognition Using Transfer Learning. IEICE Trans. Inf. Syst., 97, 2530–2532.

Stuhlsatz, A.; Meyer, C.; Eyben, F.; Zielke, T.; Meier, H.G. and Schüller, B. (2011). Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In Proceedings of the 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), Prague, Czech Republic, 22–27 May 2011.

Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schüller, B. and Zafeiriou, S. (2016). Adieu Features? End-To-End Speech Emotion Recognition Using A Deep Convolutional Recurrent Network. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016.

Truong, K.P.; van Leeuwen, D.A. and de Jong, F.M.G. (2012). Speech-based recognition of self-reported and observed emotion in a dimensional space. Speech Commun., 54, 1049–1063.

Ujwala, D., Ram Kiran, D. S., Jyothi, B., Fathima, S. S., Paruchuri, H., Koushik, Y. M. S. R. (2012). A Parametric Study on Impedance Matching of A CPW Fed T-shaped UWB Antenna. *International Journal of Soft Computing and Engineering*, *2*(2), 433-436.

Vadlamudi, S. (2015). Enabling Trustworthiness in Artificial Intelligence - A Detailed Discussion. *Engineering International*, *3*(2), 105-114. https://doi.org/10.18034/ei.v3i2.519

Vadlamudi, S. (2016). What Impact does Internet of Things have on Project Management in Project based Firms?. *Asian Business Review*, *6*(3), 179-186. https://doi.org/10.18034/abr.v6i3.520

Vlasenko, B.; Prylipko, D.; Philippou-Hübner, D. and Wendemuth, A. (2011). Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. In Proceedings of the Twelfth Annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011; pp. 1577–1580.

Williams, C.E. and Stevens, K.N. (1972). Emotions and Speech: Some Acoustical Correlates. J. Acoust. Soc. Am., 52, 1238–1250.

Wöllmer, M.; Kaiser, M.; Eyben, F.; Schüller, B. and Rigoll, G. (2013). LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. Image Vis. Comput., 31, 153–163.

--0--