Speech emotion recognition methods: A literature review

Cite as: AIP Conference Proceedings **1891**, 020105 (2017); https://doi.org/10.1063/1.5005438
Published Online: 03 October 2017

Babak Basharirad, and Mohammadreza Moradhaseli

Speech Emotion Recognition Methods: A Literature Review

Babak Basharirad [1, a)] and Mohammadreza Moradhaseli [1, b)]

[1]*School of Computing*
Asia Pacific University of Technology and Innovation (APU)
Technology Park Malaysia, Bukit Jalil, Kuala Lumpur, Malaysia

[a)] babak.basharirad@apu.edu.my
[b)] moradhaseli@outlook.com

Abstract. Recently, attention of the emotional speech signals research has been boosted in human machine interfaces due to availability of high computation capability. There are many systems proposed in the literature to identify the emotional state through speech. Selection of suitable feature sets, design of a proper classifications methods and prepare an appropriate dataset are the main key issues of speech emotion recognition systems. This paper critically analyzed the current available approaches of speech emotion recognition methods based on the three evaluating parameters (feature set, classification of features, accurately usage). In addition, this paper also evaluates the performance and limitations of available methods. Furthermore, it highlights the current promising direction for improvement of speech emotion recognition systems.

INTRODUCTION

Speech is the fast and best normal way of communicating amongst human. This reality motivate many researchers to consider speech signal as a quick and effective process to interact between computer and human. It means the computer should have enough knowledge to identify human voice and speech. Although, there is a significant improvement in speech recognition but still researcher are away from natural interplay between computer and human, since computer is not capable of understanding human emotional state. The recognition of emotional speech aims to recognize the emotional condition of individual utterer by applying his/her voice automatically. Speech emotion recognition is mostly beneficial for applications, which need human-computer interaction such as speech synthesis, customer service, education, forensics and medical analysis [1].

Recognizing of emotional conditions in speech signals are so challengeable area for several reason. First issue of all speech emotional methods is selecting the best features, which is powerful enough to distinguish between different emotions. The presence of various language, accent, sentences, speaking style, speakers also add another difficulty because these characteristics directly change most of the extracted features include pitch, energy [2]. Furthermore, it is possible to have a more than one specific emotion at the same in the same speech signal, each emotion correlate with a different part of speech signals. Therefore, defines the boundaries between parts of emotion in very challenging task. The majority of works are concentrated on monolingual emotion recognition, and making a presumption that there are no cultural diversity between utterers. However, the multi-lingual emotion classification process have been considered in some research [3].

Speech Database

In this survey different speech database are utilized to validate the proposed methods in speech emotion recognition. Among all dataset Berlin [4] and AIBO are most common used. Burkhardt et al. [4] was recorded by actors in German language. The place of record was Department of Technical Acoustics of Technical University Berlin. 5 male and 5 female German actor have participated in providing the dataset by reading one of the chosen

The 2nd International Conference on Applied Science and Technology 2017 (ICAST'17)
AIP Conf. Proc. 1891, 020105-1–020105-7; https://doi.org/10.1063/1.5005438
Published by AIP Publishing. 978-0-7354-1573-7/$30.00

sentences. Different recorded emotion are anger, fear, neutral, disgust, happiness and sadness. Another emotional database names Aibo [5] was collected in the real conditions by interacting and playing of fifty-one children with the Sony's robot Aibo that govern by human operator to extract the children's spoken speech. In AIBO five collected emotions are positive, neutral, angry, rest and emphatic.

Feature Extraction

During these years, various types of features have been proposed such as mel-frequency cepstral coefficients (MFCCs) [6] [7], and prosodic features [5] [6], linear predictive cepstral coefficients (LPCCs) [8], and perceptual linear predictive coefficients (PLPs) to achieves better consequence in human emotion recognition. Recently, there has been attention in using substitute feature extraction from an auditory-inspired long-term spectro-temporal by utilizing a modulation filterbank and an auditory filterbank for speech decomposition [9]. Figure 1 presents El Ayadi et al. [6] grouped different speech feature into 4 different groups include continuous, qualitative, spectral, and TEO (Teager energy operator)-based features. Although by selecting these features may reached a good performance but still there is some limitation which forced researchers to make their own feature sets because most of the standard features are based on short-term analysis and speech signal in a non-stationary signals.

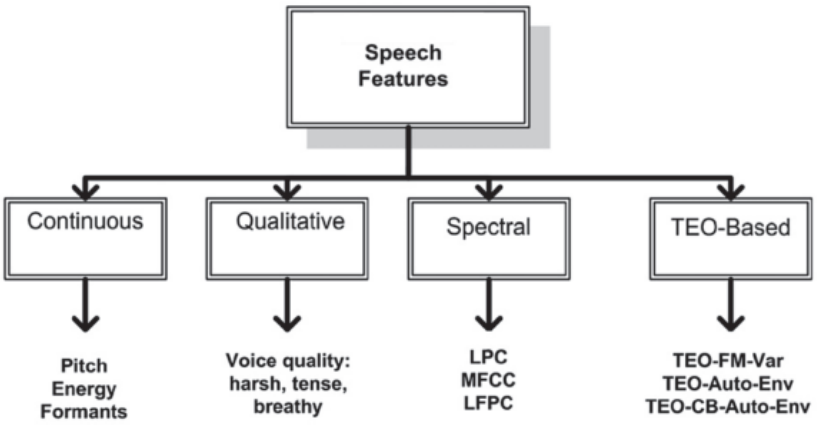

FIGURE 1. Feature Categories [1]

Classification Approaches

For modeling the emotional states, there are different classification methods utilized to create proper classifier such as support vector machine (SVM) [10][11], hidden Markov models (HMM) [12], neural network [13], K-nearest neighbor [14] and Gaussian mixture model (GMM) [6]. Conversely, a standard level of classifier may not achieve on very emotional statuses. For example ranking SVM approach cannot leads to considerable improvements in recognition of emotion compare to combination of SVM with radial basis function (RBF) [9]. Some hybrid/fusion-based methods [15] [16] achieve high recognition rate compare to individual approaches.

LITERATURE REVIEW

Over the last years, an excessive investigation has been completed to recognize emotions by using speech statistics. Cao et al. [10] proposed a ranking SVM method for synthesize information about emotion recognition to solve the problem of binary classification. This ranking method, instruct SVM algorithms for particular emotions, treating data from every utterer as a distinct query then mixed all predictions from rankers to apply multi-class prediction. Ranking SVM achieves two advantage, first, for training and testing steps in speaker- independent it obtains speaker specific data. Second, it considers the intuition that each speaker may express mixed of emotion to recognize the dominant emotion. Ranking approaches achieves substantial gain in terms of accuracy compare to conventional SVM in two public datasets of acted emotional speech, Berlin and LDC. In both acted data and the spontaneous data, which comprises neutral intense emotional utterances, ranking-based SVM achieved higher

accuracy in recognizing emotional utterances than conventional SVM methods. Unweight average (UA) or Balance accuracy achieved 44.4% [10].

Chen et al. [13] aimed to improve speech emotion recognition in speaker-independent with three level speech emotion recognition method. This method classify different emotions from coarse to fine then select appropriate feature by using Fisher rate. The output of Fisher rate is an input parameters for multi- level SVM based classifier. Furthermore principal component analysis (PCA) and artificial neural network (ANN) are employed to reduce the dimensionality and classification of four comparative experiments, respectively. Four comparative experiments include Fisher + SVM, PCA + SVM, Fisher + ANN and PCA + ANN. Consequence indicates in dimension reduction Fisher is better than PCA and for classification, SVM is more expansible than ANN for emotion recognition in speaker independent is. The recognition rates for three level are 86.5%, 68.5% and 50.2% separately in Beihang university database of emotional speech (BHUDES) [13].

Nwe et al. [12] proposed a new system for emotion classification of utterance signals. The system employed a short time log frequency power coefficients (LFPC) and discrete HMM to characterize the speech signals and classifier respectively. This method classified the emotion into six different categories then used the private dataset to train and test the new system. In order to evaluate the performance of the proposed method, LFPC is compared with the mel-frequency Cepstral coefficients (MFCC) and linear prediction Cepstral coefficients (LPCC). Result demonstrate the average and best classification accuracy achieved 78% and 96% respectively. Furthermore, results expose that LFPC is a better option as feature for emotion classification than the standard features [12].

Wu et al. [9] proposed a new modulation spectral features (MSFs) human speech emotion recognition. Appropriate feature extracted from an auditory-inspired long-term spectro-temporal by utilizing a modulation filterbank and an auditory filterbank for speech decomposition. This method obtained acoustic frequency and temporal modulation frequency components for convey important data which is missing from traditional short-term spectral features. For classification process, SVM with radial basis function (RBF) are adopted. Berlin and Vera am Mittag (VAM) are employed to evaluate MSFs. In experimental result, the MSFs display capable performance in comparison with MFCC and perceptual linear prediction coefficients (PLPC). When MSFs utilized augment prosodic features, there is a considerable improvement in performance of recognition. Furthermore overall recognition rate of 91.6% is achieved for classification [9].

Rong et al. [15] presented an ensemble random forest to trees (ERFTrees) method with a high number of features for emotion recognition without referring any language or linguistic information remains an unclosed problem. This method is applied on small size of data with high number of features. In order to evaluate the proposed method an experiment results on a Chinese emotional speech dataset designates, this method achieved improvement on emotion recognition rate. Furthermore, ERFTrees performs better than popular dimension reduction methods such as PCA and multi-dimensional scaling (MDS) and recently developed ISOMap. The best accuracy with 16 features for female dataset achieved the maximum correct rate of 82.54%, while the worst is only 16% on 84 features with natural data set.

Wu et al. [17] proposed a fusion-based method for speech emotion recognition by employing multiple classifier and acoustic-prosodic (AP) features and semantic labels (SLs). In this fusion method, first AP features are extracted then three different types of base-level classifier include GMMs, SVMs, MLP and Meta decision tree (MDT) are used. The maximum entropy model (MaxEnt) in the semantic labels method are applied. MaxEnt modeled the association between emotion association rules (EARs) and emotion states in emotion recognition. In the final state to define the emotion recognition outcome, the integrated consequence from the SL-based and AS-based are utilized. The experimental result on private dataset shows the performance based on MDT archives 80%, SL-based recognition archives 80.92, and mixture of AP and SL archives 83.55%.

Narayanan [18] proposed domain-specific emotion recognition by utilizing speech signals from call center application. Detecting negative and non-negative emotion (e.g. anger and happy) are the main focus of this research. Different types of information include acoustic, lexical, and discourse are used for emotion recognition. In addition, information-theoretic contents of emotional salience is presented to obtain data at emotion information at the language level. Both k-NN and linear discriminant classifier are used to work with different types of features. Experimental result confirms that the best results are achieved by combination of acoustic and language data. Outcomes demonstrates by combining three information source instead of one source, classification accuracy increases by 40.7% for males and 36.4% for females. Compare to pervious work improvement range in accuracy is from 1.4% to 6.75% for male and 0.75% to 3.96% for female.

Yang & Lugger [19] presented a novel set of harmony features for speech emotion recognition. These features are relying on psychoacoustic perception from music theory. First, beginning from predicted pitch of a speech signals, then computing spherical autocorrelation of pitch histogram. It calculate the incidence of dissimilar two-

pitch duration, which cause a harmonic or inharmonic impression. In Classification step, Bayesian classifier plays an important rule with a Gaussian class-conditional likelihood. Experimental result in Berlin emotion database by using harmony features indicate an improvement in recognition performance. Recognition rate improved by 2% in average [19].

Albornoz et al. [16] investigate a new spectral feature in order to determine emotions and to characterize groups. In this study based on acoustic features and a novel hierarchical classifier, emotions are grouped. Different classifier such as HMM, GMM and MLP have been evaluated with distinct configuration and input features to design a novel hierarchical techniques for classification of emotions. The innovation of the proposed method is two things, first the election of foremost performing features and second is employing of foremost class-wise classification performance of total features same as the classifier. Experimental result in Berlin dataset demonstrates the hierarchical approach achieves the better performance compare to best standard classifier, with decuple cross-validation. For example, performance of standard HMM method reached 68.57% and the hierarchical model reached 71.75% [16].

Lee et al. [11] represent a hierarchical computational structure to identify emotions. This method via following layers of binary classifications, maps input speech signal in one of the corresponding emotion classes. The main concept of different level in tree is to solve the classification task in easiest way to diminish error propagation. AIBO and USC IEMOCAP databases are employed to evaluate the classification method. Over the baseline SVM, the absolute result improve the accuracy archives an absolute improvement of 72.44%- 89.58%. The consequence proves the reported hierarchical method is efficient for classifying emotional speech in various databases [11].

Lee et al. [20] proposed hierarchical structure for binary decision tree in emotion recognition fields. This method concentrate on locates the simpler classification obstacle at top level of tree to diminish agglomeration of error. This structural method also maps input speech data into one the emotion classes via following layer of binary classification. The result on AIBO database shows 3.3% absolute over baseline model in [21], which archives 70.1% and 65.1% for two-class and five-class problem respectively. As an alternative solution instead of out- putting hard labels at every step, measuring of probability as a soft label can makes the powerful framework modeling. Bayesian Logistic Regression and SVM were employed as a binary classifier. Outcome were not significantly higher than Bjorn et al. [20, 21].

Yeh et al. [14] proposed a segment based method for recognition of emotion in Mandarin speech. This approach is contain the following process. First, define the k parameter in weighted discrete k-NN classifier, the experimental testing of different k shows the best performance for k-NN is when k sets to 10. For selecting the foremost feature set, sequential forward selection (SFS) and sequential backward selection (SBS) are employed. SFS and SBS improves feature accuracy to 84% and 82% respectively. The highest accuracy in segment-based method achieves 86%. The experimental result build on private corpus by inviting 18 males and 16 females. It is essential to gather more expressive speech to explore the extensive emotional investigation, in the future [14].

Dai et al. [7] proposed a computational approach for recognition of emotion and analysis the specifications of emotion in voiced social media such as Wechat. This approach approximate the mixed emotion and dynamic fluctuations in position- arousal-dominance (PAD) by extracting 25 acoustic features of speech signals and employing trained least squares-support vector regression (LV-SVR) model as well. The experimental results demonstrates the recognition rate for different emotion are different and the average rate of recognition achieves 82.43%, which is the best existing result by similar examination [7].

El Ayadi et al. [6] proposed a Gaussian mixture vector autoregressive (GMVAR) approach, which is mixture of GMM with vector autoregressive for classification problem of speech emotion recognition. The key idea of GMVAR is its capability to multi-modality in their dissemination and design the dependency between speech feature set. Berlin emotional dataset was used for evaluation of GMVAR. The experimental result shows classification accuracy achieves 76% when for HMM reached 71%, for k-NN 67% and 55% for feed-forward neural networks. The advantage of this method better differentiation amongst high and low arousal with neutral emotions compare to HMM [6].

Arias et al. [22] proposed a novel shape based method by using neutral model to recognize emotional salience in the basic frequency. The new method supported by functional data analysis (FDA) that target to obtain the natural changeability of F0 contours. For a certain F0 contour, PCA are calculated to use as feature for speech emotion recognition. The empirical result indicate the proposed approach accuracy achieved 75.8% in binary classification. It means 6.2% higher than trained benchmark system with over-all F0 statics. The approach is evaluated by the SEMAINE dataset. The results designate that to recognize speech emotion, utilizing the shape-based method in real application can be effective [22].

Grimm et al. [23] proposed a multi-dimensional model by utilizing emotion primitives for speech emotion recognition. Three dimension were made by composing of three different value of emotion primitives, which is

called valence, activation, and dominance. The value of these factors assumed to be in the range of [-1, +1]. A text-free, image-based method was introduced to assess the emotion primitives and achieves best inter-evaluator agreement. For extracting acoustic feature such as energy, pith and spectral specifications, both fuzzy logic and rule-based estimator are employed. The approached are validate by testing two EMA and VAM datasets, which are acted emotion and spontaneous speech emotion. Both dataset are recorded form talk-show in German TV. Finally, for mapping the emotion primitives to certain emotion category, k-NN was employed as a classifier. K-NN achieves total recognition rate up to 83.5% [23].

COMPARISON STUDY AND FEATURE DIRECTION

Generally, analysis on speech emotion recognition aims to increase the recognition rate as well as accuracy. Table-1 presents the current available methods, which targeted the speech emotion recognition systems, and these are evaluated with its classifier, features set, and recognition rate and on different dataset levels. As table-1 depicts that SVM has many motivating properties such as easy to fulfilment because of mathematical basis compare to simple classifier such as quadratic discriminant analysis (QDC) [22]. Normally, SVM is employed as alone and also with combination of other classifier such as ANN and RBF to reduce the dimensionality. It shows most accuracy on according to table-2, where it uses the log frequency power coefficient (LEPC) and MFCC features set and achieve more than 95% in best case. The main limitation author just tested on private dataset. It should be tested with Berlin and VAM dataset or some other to verify the accuracy. Besides SVM, k-nearest neighbor (k-NN) is one of the well-known classifier in speech emotion recognition. Yeh et al. [14] have proven in the best situation can achieve 86% of recognition rate using k-NN method. Nwe et al. [12] by using HMM and adopting LFPC as feature parameters achieves the highest recognition rate in this research, which is 96% in the best situation, and his average rate is 78%.

TABLE 1. Different Types of Features, Classifier and Dataset in Current Speech Emotional Recognition System

Ref	Types of classifier	Types of features	Recognition Rate	Type of Dataset	Methods
[10]	SVM	Prosodic and spectral features	44.4%	Berlin & LDC & FAU Aibo dataset	Ranking SVM
[13]	SVM & ANN	Energy, ZCR, pitch, SC, spectrum cut-off frequency, correlation density (Cd), fractal dimension, MFF.	86.5%, 68.5% and 50.2% for different level	Beihang University Database of Emotional Speech (BHUDES)	Multi-level SVM classifier & ANN to reduce dimensionality
[12]	HMM	Log frequency power coefficients (LFPC), MFCC	Average and best result 78% and 96% respectively	Two private speech dataset	Discrete HMM and LFPC to characterize speech signal
[9]	SVM & RBF	Modulation spectral features (MSFs)	91.6%	Berlin & VAM	Modulation filterbank & auditory filterbank for speech decomposition, SVM & RBF for classification
[15]	Decision tree & random forest	Linguistic , spectral-related , contour-related, tone-based and /or vowel-related features	Best 82.54% & worst 16%	Private series of Chinese emotional speech dataset	ensemble random forest to trees (ERFTrees) method with a high number of features
[17]	GMM, SVM, MLP and MDT(Meta decision tree)	Prosodic information and semantic labels (SL), acoustic and prosodic features	MDT 80%, SL-based 80.92%, mixture of AP & SL 83.55%	Private database	Fusion method based on AP & SL and multiple classifier by maximum entropy model (MaxEnt)
[18]	k-NN & linear discriminate	Fundamental frequency (F0), energy, duration, and the first and formant	40.7% for males & 36.4% for females	Private speech database from call center	domain-specific emotion recognition by k-NN and linear discriminate classifier
[19]	Bayesian learning framework	Energy, pitch statistics, duration, formant, and zero-crossing rate (ZCR)	Best for sadness 87.6% and overall 2% improvement	Berlin emotion dataset	Harmony features with Bayesian classifier with Gaussian class conditional likelihood
[16]	HMM, GMM, MLP and hierarchical model	Mean of the log-spectrum (MLS), MFCCs and prosodic features	HMM 68.57, Hierarchical model 71.75	Berlin dataset	Spectral characteristics of signals are used in order to group emotions based on acoustic rather than psychological considerations.
[11]	SVM	Zero crossing rate, root mean square energy, pitch, harmonics-to-noise ratio, and MFF	72.44% - 89.58%	AIBO and USC IEMOCAP dataset	Hierarchical computational structure to maps an input speech utterance into one of the multiple emotion classes
[20]	Bayesian Logistic Regression, SVM	large-margin feature	70.1% & 65.1% for two and five class	AIBO dataset Private Mandarin	Hierarchical structure for binary decision tree
[14]	k-NN	Jitter, shimmer, formants, LPC, LPCC, MFCC, LFPC, PLP, and Rasta-PLP, SFS and SBS	Best 86%	Chinese emotional speech corpus We invited 18 males and 16 females.	Segment based method by employing k-NN, SFS (sequential forward selection), SBS (sequential backward selection)
[7]	LS-SVR	Short-time Energy , Pitch , Short-time zero crossing rate, First and second Formant	82.43%	Samples of 180 chats from the historical data on Wechat.	Computational approach by analyzing approximate the mixed emotion and dynamic fluctuations in position- arousal-dominance (PAD) employing trained least squares-support vector regression (LV-SVR) model as well.
[6]	G GMVAR	Mel-frequency cepstrum coefficient (MFCC),	76%	Berlin emotional speech database,	Gaussian mixture vector autoregressive (GMVAR) is a mixture of GMM with vector autoregressive for classification.
[22]	Binary classifier & QDC	Prosodic features such as energy contour and duration	75.8%	SEMAINE databases	Shape based method by using functional data analysis to obtain natural changeability of F0
[23]	k-NN	Acoustic features such as pitch, energy, speaking rate and spectral characteristics.	83.5%	EMA and VAM dataset	A multi-dimensional model by utilizing emotion primitives. Three dimension were made by composing of three different value of emotion primitives, which is called valence, activation, and dominance.

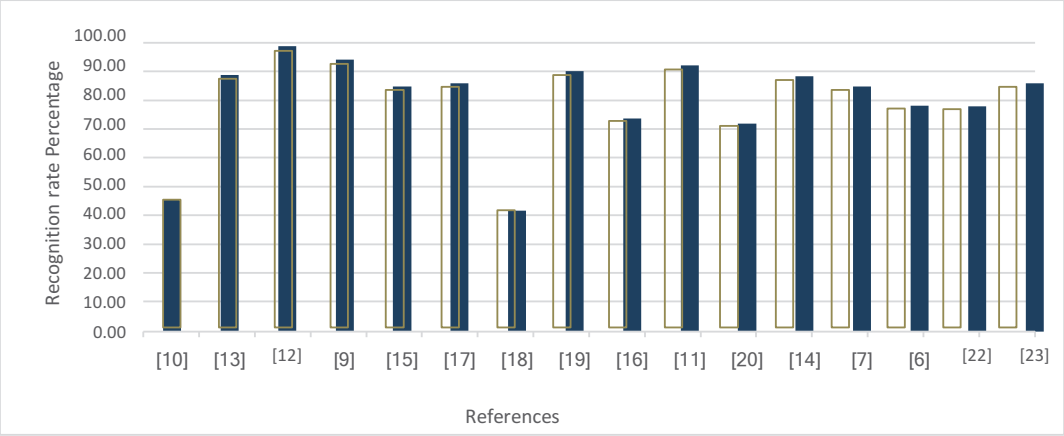

FIGURE 2. Comparision of Speech Emotion Recognition in Terms of Recognition rate

CONCLUSION

In this paper, we reviewed and discussed various speeches emotional recognition systems based approaches. We also compare its performance in terms of classifier, features, recognition rate, and datasets. Well-design classifiers have obtain high classification accuracies between different types of emotions. In this survey HMM with adopting short time LFPC as a feature proves a good accuracy on different levels in the chart. Most of the current research concentrate on investigating different features and their correlation with emotional state in spoken speech. In this fact some researcher develop their own feature like MLS to achieve high performance in recognition rate. The majority of the current datasets are not capable for evaluation of speech emotion recognition. In most of them, it is hard even for human to specify different emotion of certain collected utterances; e.g. the human recognition accuracy was 80% for Berlin [4]. In conclusion, there are only limited studies that considered applying multiple classifier to speech emotion recognition [16] [17]. We consider multiple classifier methods (MCM) as a further research direction, which has to be, explore in future.

REFERENCES

1. M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011.
2. R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," Pers. Soc. Psychol, vol. 70, no. 3, pp. 572–587, 1996.
3. V. Hozjan and Z. Kačič, "Context-Independent Multilingual Emotion Recognition from Speech Signals," Int. J. Speech Technol., vol. 6, no. 3, pp. 311–320, 2003.
4. F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech.," in Interspeech, 2005, vol. 5, pp. 1517–1520.
5. A. Schuller, B. , Steid, S. l, and Batliner, "The interspeech 2009emotion challengee," Interspeech, pp. 312–315, 2009.
6. M. M. H. El Ayadi, M. S. Kamel, and F. Karray, "Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models," in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 2007, vol. 4, pp. IV–957–IV–960.
7. W. Dai, D. Han, Y. Dai, and D. Xu, "Emotion Recognition and Affective Computing on Vocal Social Media," Inf. Manag., Feb. 2015.
8. B. S. Atal, "Effectiveness of liner prediction characteristics of the speech wave for automatic speaker speech wave for automatic speaker identification and verification," Acoust. Soc. Am., vol. 55, no. 6, pp. 1304–1312, 2005.
9. S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.
10. H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.
11. C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," Speech Commun., vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011.
12. T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.
13. L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.
14. J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen, "Segment-based emotion recognition from continuous Mandarin Chinese speech," Comput. Human Behav., vol. 27, no. 5, pp. 1545–1552, Sep. 2011.
15. J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009.
16. E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," Comput. Speech Lang., vol. 25, no. 3, pp. 556–570, Jul. 2011.
17. C.-H. Wu and W.-B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 10–21, Jan. 2011.
18. S. S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2005.
19. B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," Signal Processing, vol. 90, no. 5, pp. 1415–1423, May 2010.
20. C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," Interspeech, vol. 53, pp. 320–323, 2009.
21. S. Bjorn, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," 2009. .
22. J. P. Arias, C. Busso, and N. B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," Comput. Speech Lang., vol. 28, no. 1, pp. 278–294, Jan. 2014.
23. M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," Speech Commun., vol. 49, no. 10–11, pp. 787–800, Oct. 2007.