2016

BigEAR: Inferring the Ambient and Emotional Correlates from Smartphone-based Acoustic Big Data

Harishchandra Dubey

Matthias R. Mehl

See next page for additional authors

Authors

Harishchandra Dubey, Matthias R. Mehl, and Kunal Mankodiya

BigEAR: Inferring the Ambient and Emotional Correlates from Smartphone-based Acoustic Big Data

Harishchandra Dubey*, Matthias R. Mehl†, Kunal Mankodiya*

*Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA
†Department of Psychology, University of Arizona, Tucson, AZ, USA

dubey@ele.uri.edu, mehl@email.arizona.edu, kunalm@uri.edu

Abstract—This paper presents a novel BigEAR big data framework that employs psychological audio processing chain (PAPC) to process smartphone-based acoustic big data collected when the user performs social conversations in naturalistic scenarios. The overarching goal of BigEAR is to identify moods of the wearer from various activities such as laughing, singing, crying, arguing, and sighing. These annotations are based on ground truth relevant for psychologists who intend to monitor/infer the social context of individuals coping with breast cancer. We pursued a case study on couples coping with breast cancer to know how the conversations affect emotional and social well being. In the state-of-the-art methods, psychologists and their team have to hear the audio recordings for making these inferences by subjective evaluations that not only are time-consuming and costly, but also demand manual data coding for thousands of audio files. The BigEAR framework automates the audio analysis. We computed the accuracy of BigEAR with respect to the ground truth obtained from a human rater. Our approach yielded overall average accuracy of 88.76% on real-world data from couples coping with breast cancer.

I. INTRODUCTION

Big data have revolutionize all areas of everyday life such as healthcare, psychology, psychiatry [1]–[3]. Automated speech recognition has become a part of our life that involves constant interactions with smartphones. However, speech emotion recognition is a challenging problem, especially when acoustic big data is filling up the cloud storage at a rapid rate. For example, our previous work, known as iEAR [4], recorded a monumental acoustic data from iPhones that involve social and daily life conversations of couples coping with breast cancer. The physiological state of the individuals coping with breast cancer affects their emotions and intimacy that is reflected in everyday conversations. For instance, the mood of such individuals at various times in a day can be inferred from their daily audio log (DAL). Collection of audio data was automated that facilitated unbiased capture of ambient sounds

in naturalistic settings. The psychologists had an access to the audio data and also the estimated mood labels that can help in making clinical decisions about the psychological health of the participants. Recently, our team developed a novel framework named as BigEAR that employs a psychological audio processing chain (PAPC) to predict the mood of the speaker from DALs. The BigEAR application installed on the iPhone/iPad captures the audio data from ambient environment at random instance throughout the day. The snippets are of small sizes so that we do not record significant duration of a personal context. The acquired audio data contain speech, music as well as silence. The audio data contain rich information about the emotional correlates in the social conversation such as happiness, sadness or anger. The presented work makes the following contributions to the area of big data science and analytics:

- Development and validation of psychological audio processing chain (PAPC) to generate big-data-driven knowledge about the emotion correlates;
- Proposed novel perceptual features such as frequency modulation, frequency range, and sharpness for identifying the emotions from speech;
- Use of advanced tools for speaker diarization (segmentation) on psychological audio big data;
- Automating the process of emotion identification from audio data in context of couples coping with breast cancer.

II. BACKGROUND & RELATED WORKS

Identifying the surrounding environment and emotional correlates of the naturalistic conversation between individuals is an important part of the psychologists treatment regime for couples coping with breast cancer [5], [6], [7]. However, all our previous studies and similar studies by others involved manual transcription and analysis of audio by subjective hearing to decide ambient and emotional correlates for couples coping with breast cancer [8]. The increasing use of smartphones, smartwatches and wrist-bands provides a convenient interface for collection of audio data without need of any specialized hardware [3], [9]–[12]. We designed
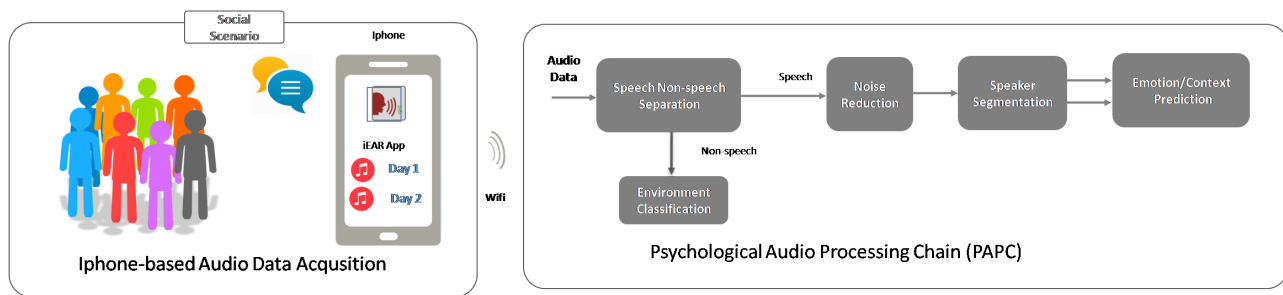
Fig. 1. A conceptual view of the BigEAR framework. The iEAR application installed on the iPhone captures the audio data from users. The acquired audio data contain speech and non-speech components. The speech data contain rich information about the emotional correlates such as happiness, sadness or neutral emotion while non-speech contain rich information about the surrounding environment such as outdoor, indoor, and TV/music. The audio data were processed by a psychological audio processing chain (PAPC) developed by our team to predict the emotion correlates and identify the surrounding environment. Collection of audio data was automated that facilitated unbiased capture of ambient sounds in naturalistic settings.

iEAR that is an iPhone app used for unobtrusive collection of ambient sound from users surroundings. There are some similar systems in literature such as authors in [13] developed a smart system for collecting the real-life experiences of peoples and store it into an adaptive multimedia bank with variable size. The system used a camera, a microphone and an accelerometer for combined collection of audio and video data when users perform interesting activities. There was a sub-block that detects the presence of interesting activity [13]. However, it required users to use a specialized hardware and wear it for data collection. In contrast, iEAR does not require any specialized hardware. We have used iEAR for audio collection in various research problems in social psychology. For example, we studied self-awareness in individuals using iEAR recorder followed by manual transcription [14], [6]. The individual beliefs regarding their behavior were compared with the inferences brought out by the team of psychologists and clinicians from the collected audio data. The conformance between two metrics quantifies the self-knowledge individual possesses about themselves. We concluded that self-knowledge was positively correlated with the individual's perception of relationship quality. Since the quality of our relationships greatly affects our social life, individual's self-knowledge can add value to their life [15]. We had devised a method for assessment of social well being and health of individual using iEAR and manual transcription. Such approaches are useful for health researchers and psychologists. iEAR allows social health scientists to hear the audio snippets from individual's personal and social life and hence cross-validate research findings against self-reports by individuals [16]. In this paper we focus on automated audio analysis-based inference on audio collected by iEAR during social interactions of couples coping with breast cancer. From the ambient audio data, we can know the mood of the couples. The moods are categorical labels that quantify the context of conversation, intensity and type of emotions and sentiment linked to the conversation. The features are usually fed to a classification systems for predicting the class of given pattern classification problem [17], [18]. At the end of this article, we will see the benefits of BigEAR framework for automated psychological
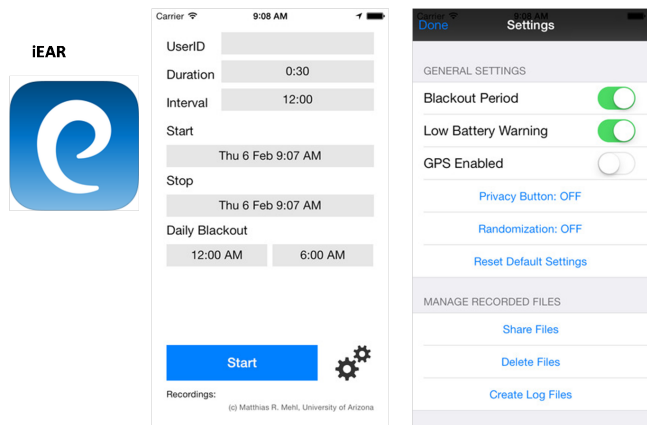

Fig. 2. The BigEAR app used for recording daily life activities of users in form of daily-audio-log (DAL). It unobtrusively records the ambient sounds from the surroundings periodically for 50 seconds duration giving out snippets of ambient sounds.

audio processing thus reducing logistics and manpower needs.

III. PSYCHOLOGICAL AUDIO PROCESSING CHAIN (PAPC)

Figure 1 shows the conceptual view of PAPC framework proposed in this paper. It is a cascade of audio processing blocks for realizing psychology-based speech processing. The audio data is digitized as a mono-channel signal sampled at 11.025 kHz sampling rate with 16-bit precision. The collected audio data inherently contains noise and significant amount of silences. It necessitated the preprocessing the audio with silence suppression and noise reduction before feeding it to feature extraction and mood classification system. For audio processing, we employed overlapping 40ms windows with skip rate of 10ms.

A. Silence Removal

The speech non-speech decision was made by the algorithm developed in [20]. It is based on the effective selection of audio-frames. The short time-windows of the speech signal are stationary (for 25-40ms windows). However, for an extended time-duration (more than 40ms), the statistics of the speech
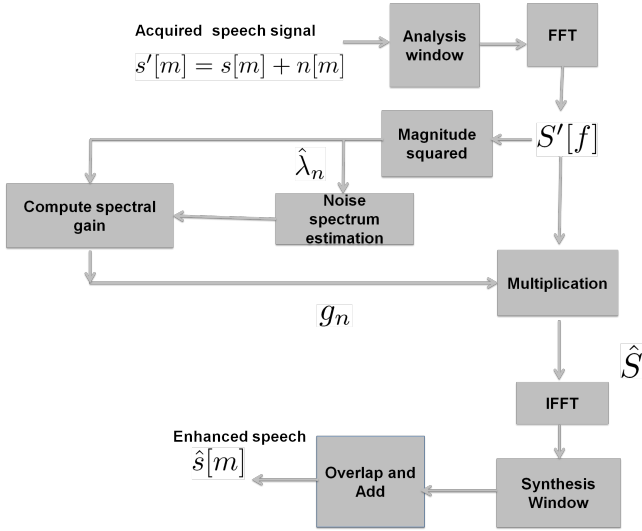
Fig. 3. Block diagram of the method used for reducing the background noise [19].
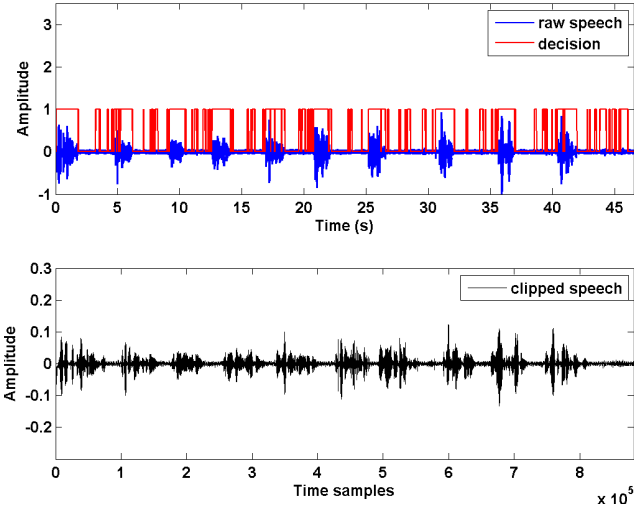

Fig. 4. The top sub-figure shows the time-domain audio signal and the corresponding speech non-speech decision. The bottom sub-figure shows the separated speech contained in audio recording [20].
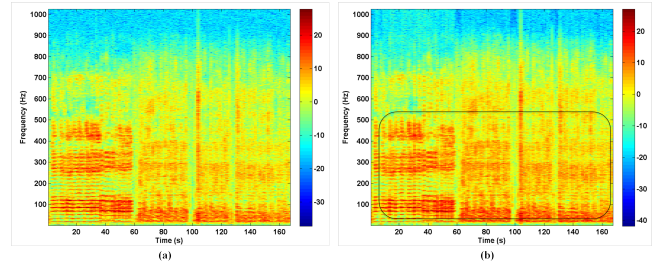

Fig. 5. The spectrogram of (a) original speech signal and (b) enhanced signal. Increase in energy of certain frequencies marked by squared block is clearly evident.

speech signal using noise estimates from minima controlled recursive averaging. Two functions were used for accounting the probability of speech presence in various sub-bands. One of these functions was based on the time frequency distribution of the a-priori SNR and was used for estimation of the speech spectrum. Second function was decided by the ratio of energy of noisy speech-segment and its minimum value within the time-window. Figure 3 shows the block diagram of the noise reduction method proposed in [21], [19]. Figure 5 shows the spectrogram of speech signal and that of corresponding enhanced signal. Energy enhancement is clearly visible in the frequency range 10-500 Hz that is marked by the square block. Let $S[n]$ and $S'[n]$ be the speech before and after noise reduction, we can derived a metric, change in signal-to-noise energy (CSNE) in dB as follows:

$$CSNE_{dB} = \frac{\sum_{n=0}^{L-1} S'[n]^2}{\sum_{n=0}^{L-1} \left(S'[n] - S[n]\right)^2} \qquad (1)$$

where L is the length (in samples) of the speech signals and n is the time-index. CSNE in dB can be used for environment classification as discussed in next sections. Usually, outdoors are accompanied by high CSNE values as compared to indoors. The TV/music gave intermediate values for CSNE.

C. Speaker Segmentation in Speech Signal

The audio signal contains voices of various speakers. Before making inference on moods of each one of the speaker such as couples, we have to separate the speaker segments in audio signal. We used the method developed in [22] for speakers diarization of enhanced speech signal after noise reduction as shown in Figure 1. Authors used Mel-frequency Cepstral Coefficient (MFCC) features for finding Fisher near-optimal linear discriminant subspace for developing an unsupervised method for speaker segmentation. The proposed approach relies on a semi-supervised version of Fisher linear discriminant analysis (FLD) and utilizes the sequential structure of speech signal. The overall accuracy of 92 % with respect to ground-truth was obtained for speaker diarization using this technique on our data. Readers are referred to [22] for more details on implementation strategy.

signals changes significantly rendering unequal relevance of the speech-frames. It necessitates the selection of effective frames on the basis of posteriori signal to noise ratio (SNR). The authors used the energy distance for measuring the relevance of speech-frames. This method was found to be effective even at low SNRs and had reduced computational complexity as compared to other available algorithms. Hence, we chose it for suppressing the silences in audio recording obtained from iEAR app. Figure 4 illustrates the speech non-speech decision for an audio signal.

B. Noise Reduction

The speech signal inevitably contain background noise. Authors in [19] developed a method for reducing non-stationary noise in audio signal. It optimized log-spectral amplitude of

D. Classification of Acoustic Environment

The background-noise can quantify the acoustic environment where social conversation take place. For example, on an average outdoors are usually more noisy as compared to indoors. In this paper, we have used change in signal-to-noise energy (CSNE) in dB and speech signal as features in a decision tree classifier for classification of environment into three categories namely, indoor, outdoor, and TV/music. These features were computed on short overlapping frames of 40 ms with 10 ms skip rate. All the frame-level features were cascaded into a feature vector and fed to a C4.5 decision-tree classifier [23]. We achieved an overall accuracy of 83.68% using these features and decision tree classifier.

IV. EXPERIMENTS

A. Smart Audio Data

The participants were recruited by the Department of Psychology in University of Arizona for collection of audio data. All the participants signed the consent form approved by the University of Arizona Institutional Review Board before collecting the psychological audio data. The individuals using iEAR were couples coping with breast cancer. The iEar app running on iPhones collected short chunks of audio data (50 seconds) at random instances throughout the day, thus, capturing the acoustic scenarios at irregularly spaced time intervals [24]. Each day accumulated 80 files on average from each participant. We used the data from 10 individuals who used iEAR app over 1 week, thus a total of 5600 audio recording, each one of 30 seconds duration. Manual transcription was obtained for each of these files along with mood labels. We divided the corpus into two parts: training set and testing set. Out of total, 5600 audio files, 4500 was used for training the classification system and rest 1100 for testing. Same division of data was used for determining the accuracy of speech activity detection.

B. Perceptual Audio Features

We designed some new features for our task and fused it with the state-of-the-art features for emotion detection from speech and music. The Munich Versatile and Fast Open-Source Audio Feature Extractor (OpenSMILE) is a feature extractor for emotion recognition. We used the default parameter set in v2.1 openSMILE for research discussed in this paper. OpenS-MILE had been successfully applied for emotion identification in music, speech and other mixed sources, [25]. Basically, the OpenSMILE features consists of sound, speech, and music features along with cross-domain features. The energy, spectral and voicing related low-level descriptors (LLDs) are used. We processed the audio files for noise reduction and suppression of silences before feeding it to openSMILE feature extractor. The features were observed to be correlated that necessitated preprocessing with DCT before feeding it into classification system. Next, we used DCT for decorrelating the features set (6373 dimensional features from OpenSMILE), thus reducing the dimension to 3000. Later we appended our curated feature set to OpenSMILE post-processed features.

1) Jitter: Jitter (J_1) quantifies cycle-to-cycle changes in the vocal period. Fundamental frequency contour was used to obtain the two features namely absolute and relative jitter [26]. J_1 is defined as the average absolute difference between consecutive periods. Mathematically, it is given as:

$$J_1 = \frac{1}{M} \sum_{j=1}^{M-1} |F_j - F_{j+1}| \tag{2}$$

where F_j is the j-th extracted vocal period and M is the number of extracted vocal periods. *Relative Jitter*, J_2, is defined as the ratio of average absolute difference between consecutive vocal periods to the average period. Mathematically, it is given by (expressed as percentage):

$$J_2 = \frac{\frac{100}{M} \sum_{j=1}^{M-1} |F_j - F_{j+1}|}{\frac{1}{M} \sum_{j=1}^{M-1} |F_j|} \tag{3}$$

2) Shimmer: It measures variations in intensity of the speech signal. It captures the instabilities in oscillating pattern of the vocal folds that influences the cycle-to-cycle changes in speech amplitude. It is defined as average absolute difference in intensities from two successive cycles. Shimmer in decibels (dB) is mathematically, given as:

$$S_{dB} = \frac{20}{M} \sum_{j=1}^{M-1} |\log 10 \frac{A_i}{A_{i+1}}| \tag{4}$$

where A_i is the maximum amplitude of speech in i-th frame and M is the number of frames.

3) Frequency Modulation: It quantifies the presence of sub-harmonics in speech signal. There are no references for its use in emotion classification. Usually, speech signals with many sub-harmonics lead to a more complicated interplay between various harmonic components making it relevant for emotion identification. Mathematically, it is given as [27]:

$$F_{mod} = \frac{\max (F_j)_{j=1}^{M} - \min (F_j)_{j=1}^{M}}{\max (F_j)_{j=1}^{M} + \min (F_j)_{j=1}^{M}} \tag{5}$$

where F_{mod} is frequency modulation, and F_j is the fundamental frequency of j-th speech frame.

4) Frequency Range: The range of frequencies is an important feature of speech signal that quantifies its quality and emotions associated with it [28]. We computed the frequency range as the difference between 5-th and 95-th percentiles. Mathematically, it becomes:

$$F_{range} = F_{95\%} - F_{5\%} \tag{6}$$

Taking 5-th and 95-th percentiles helps in eliminating the influence of outliers in estimates of fundamental frequency.

5) Harmonics to Noise Ratio: Harmonics to Noise Ratio (HNR) quantifies the noise present in the speech signal that results from incomplete closure of the vocal folds during speech production process [26]. In this paper, we have used method proposed in [29] for HNR estimation. The average and

S.No.	Moods category	Physical Interpretation
1	laugh	situation where either primary or secondary speaker is laughing
2	sing	either of speakers are singing or from TV/surroundings
3	cry	either of speakers are singing or from TV/surroundings
4	arguing	between couples
5	sigh	from either of speakers or from TV/surroundings

standard deviation of the segmental HNR values are later used for emotion prediction. Lets assume that R_{xx} is normalized autocorrelation and $l_{\max}$ is the lag (in samples) at which it is maximum, except the zero lag. Then, HNR is mathematically given by [29]:

$$HNR_{dB} = 10\log 10 \left(\frac{R_{xx}(l_{\max})}{1 - R_{xx}(l_{\max})} \right) \quad (7)$$

6) Spectral Centroid: It is the *center of mass* of spectrum. It measure the *brightness* of an audio signal. Spectral centroid of a spectrum-segment is given by average values of frequency weighted by amplitudes, divided by the sum of amplitudes. Mathematically, we have

$$SC = \frac{\sum_{n=1}^{N} kF[k]}{\sum_{n=1}^{N} F[k]} \quad (8)$$

where SC is the spectral centroid, and $F[k]$ is amplitude of k-th frequency bin of discrete Fourier transform of speech signal. Average spectral centroid for complete audio signal was used for emotion identification [30].

7) Spectral Flux: It quantifies the rate of change in power spectrum of speech signal. It is calculated by comparing the normalized power spectrum of a speech-frame with that of other frames. It determines the timbre of speech signal. Spectral flux had been previously used for emotion recognition in music [31].

8) Spectral Entropy: Spectral Entropy was used for emotion recognition in speech [32]. It is given by:

$$SE = \frac{-\sum P_j log(P_j)}{log(M)} \quad (9)$$

where SE is the spectral entropy, P_j is the power of j-th frequency-bin and M is the number of frequency-bins. Here, $\sum P_k = 1$ as the spectrum is normalized before computing the spectral entropy.

9) Spectral Flatness: It measures the flatness of speech power spectrum. It quantifies how similar the spectrum is to that of a noise-like signal or a tonal signal. Spectral Flatness (SF) of white noise is 1 as it has constant power spectral density. A pure sinusoidal tone has SF close to zero showing the high concentration of power at a fixed frequency. Mathematically, SF is ratio of geometric mean of power spectrum to its average value [33].

10) Sharpness: Sharpness is a mathematical function that quantifies the sensory pleasantness of a speech signal. High sharpness implies low pleasantness. It value depends on the spectral envelope of signal, amplitude level and its bandwidth. The unit of sharpness is acum (Latin expression). The reference sound producing 1 acum is a narrowband noise, one critical band wide with 1 kHz center frequency at 60 dB intensity level [34]. Sharpness, S is mathematically defined as

$$S = 0.11 \frac{\sum_0^{24Bark} L_0 \cdot g(z) \cdot z \cdot dz}{\sum_0^{24Bark} L_0 \cdot dz} acum \quad (10)$$

However, its numerator is weighted average of specific loudness (L_0) over the critical band rates. The weighting function, $g(z)$, depends on critical band rates. The $g(z)$ could be interpreted as the mathematical model for the sensation of sharpness.

V. RESULTS & DISCUSSIONS

The 11 dimensional curated features consisting of absolute and relative jitter, shimmer, frequency modulation, frequency range, Harmonic-to-noise ratio,spectral centroid, spectral flux, spectral entropy, spectral flatness and sharpness are cascaded with 3000 dimensional DCT-processed OpenSMILE features. These final features of dimension 3011 are later fed to a deep neural network(DNN) with 4 hidden layer with 2048, 2048, 1024 and 1024 nodes respectively. The input node had 3011 dimension while output node had 5 nodes. We used log-sigmoid non-linearity in hidden layers and soft-max linearity of output layer. The cross-entropy between ground-truth and output was used as objective function in training the DNN. Deep Neural Networks (DNNs) are large scale version of feed-forward neural networks that have been successfully validated for learning the complex functional relations. Use of DNN for detecting emotions from speech and music have been well accounted in literature [35]. Each feature dimension was normalized to have zero mean and unit variance before feeding it to DNN. Dropout strategy was used to train the DNN as it avoids over-fitting that can happen in high dimensional neural network with hybrid features from music and speech [36]. We used kaldi toolkit for realizing DNN training and testing as it is popular for training deep neural networks especially in speech recognition community [37]. The overall average accuracy of DNN classifier with complete feature set comes out to be 88.76% that is a competitive accuracy owing to naturalistic noisy conditions where the data recording takes place. It was observed from the collected audio data that the cancer related discussion were more common with couples as

compared to friends and family. The discussion were more informational than emotional. In addition, the involvement of couples in discussions lead to be adjustment in cancer patients. The main benefit of using BigEAR is that it provides objective criterion for such studies and evaluation.

VI. CONCLUSIONS

This paper developed the BigEAR framework was validated to be a unobtrusive smartphone-based system that could identify the emotion and ambient correlates during social conversations in naturalistic scenarios. The case study involved couples coping with breast cancer. The BigEAR framework is powered by psychological audio processing chain (PAPC) for inferring the environment where the conversations take place and may contain mood or emotional information of the speakers. The advantage of BigEAR framework is that the psychologists no longer analyze the increasing amount of acoustic big data that demand listening to each and every audio file and classify the moods into categories. The downside is that the BigEAR is just 88.76% accurate as compared to ground-truth labels from human listeners. However, this accuracy level is in an acceptable range for clinicians and psychologists since BigEAR offers automated audio analysis as opposed to human ratings that are time consuming, costly, and complicated w.r.t. manpower management.

REFERENCES

[1] V. Mayer-Schönberger and K. Cukier, *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[2] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[3] H. Dubey, J. Yang, N. Constant, A. M. Amiri, Q. Yang, and K. Makodiya, "Fog data: Enhancing telehealth big data through fog computing," in *Proceedings of the ASE BigData & SocialInformatics 2015*. ACM, 2015, p. 14.

[4] M. R. Mehl and S. E. Holleran, "An empirical analysis of the obtrusiveness of and participants' compliance with the electronically activated recorder (ear)," *European Journal of Psychological Assessment*, vol. 23, no. 4, pp. 248–257, 2007.

[5] M. Mets, "Naturalistically observed social support and optimism in couples coping with breast cancer," 2012.

[6] M. L. Robbins, A. M. López, K. L. Weihs, and M. R. Mehl, "Cancer conversations in context: Naturalistic observation of couples coping with breast cancer." *Journal of Family Psychology*, vol. 28, no. 3, p. 380, 2014.

[7] M. L. Martino, R. Onorato, and M. F. Freda, "Linguistic markers of processing trauma experience in womens written narratives during different breast cancer phases: Implications for clinical interventions," *Europes Journal of Psychology*, vol. 11, no. 4, pp. 651–663, 2015.

[8] K. Cheng, R. D. Devi, W. Wong, and C. Koh, "Perceived symptoms and the supportive care needs of breast cancer survivors six months to five years post-treatment period," *European Journal of Oncology Nursing*, vol. 18, no. 1, pp. 3–9, 2014.

[9] H. Dubey, J. C. Goldberg, M. Abtahi, L. Mahler, and K. Mankodiya, "Echowear: smartwatch technology for voice and speech treatments of patients with parkinson's disease," in *Proceedings of the conference on Wireless Health*. ACM, 2015, p. 15.

[10] H. Dubey, J. C. Goldberg, K. Makodiya, and L. Mahler, "A multi-smartwatch system for assessing speech characteristics of people with dysarthria in group settings," in *Proceedings e-Health Networking, Applications and Services (Healthcom), 2015 IEEE 17th International Conference on, Boston, USA*, 2015.

[11] L. Mahler, H. Dubey, J. C. Goldberg, and K. Mankodiya, "Use of smartwatch technology for people with dysarthria," in *accepted in Proceedings of the conference on Motor Speech, Newport Beach, California.* Madonna Rehabilitation Hospital, 2016.

[12] A. Monteiro, H. Dubey, L. Mahler, Q. Yang, and K. Mankodiya, "Fit: A fog computing device for speech teletreatments," *2nd IEEE International Conference on Smart Computing (SMARTCOMP 2016), At Missouri, USA*, 2016.

[13] M. Blum, A. S. Pentland *et al.*, "Insense: Interest-based life logging," *IEEE MultiMedia*, no. 4, pp. 40–48, 2006.

[14] M. R. Mehl, M. L. Robbins, and F. große Deters, "Naturalistic observation of health-relevant social processes: The electronically activated recorder (ear) methodology in psychosomatics," *Psychosomatic Medicine*, vol. 74, no. 4, p. 410, 2012.

[15] E. R. Tenney, S. Vazire, and M. R. Mehl, "This examined life: The upside of self-knowledge for interpersonal relationships," *PloS one*, vol. 8, no. 7, p. e69605, 2013.

[16] M. R. Mehl, "Eavesdropping on health: a naturalistic observation approach for social health research," *Social and personality psychology compass*, vol. 1, no. 1, pp. 359–380, 2007.

[17] H. Dubey and A. K. Tiwari, "Blind modulation classification based on mlp and pnn," in *Engineering and Systems (SCES), 2012 Students Conference on*. IEEE, 2012, pp. 1–6.

[18] H. Dubey, A. Tiwari, P. Ray, S. Mohanty, and N. Kishor, "A novel fault classification scheme based on least square svm," in *Engineering and Systems (SCES), 2012 Students Conference on*. IEEE, 2012, pp. 1–5.

[19] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 466–475, 2003.

[20] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 798–807, 2010.

[21] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[22] T. Giannakopoulos and S. Petridis, "Fisher linear semi-discriminant analysis for speaker diarization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 1913–1922, 2012.

[23] J. Quinlan, "C4. 5: Program for machine learning morgan kaufmann," *San Mateo, CA, USA*, 1993.

[24] M. Mehl, "https://itunes.apple.com/us/app/iear/id451939501?mt=8," accessed: 2016-01-216.

[25] F. Eyben and B. Schuller, "Music classification with the munich opensmile toolkit," in *Proceedings of the Annual Meeting of the MIREX 2010 community as part of the 11th International Conference on Music Information Retrieval (ISMIR)*, 2010.

[26] A. Tsanas, "Accurate telemonitoring of parkinsons disease symptom severity using nonlinear speech signal processing and statistical machine learning," Ph.D. dissertation, University of Oxford, 2012.

[27] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio," pp. 676–679, 2000.

[28] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[29] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[30] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 617–620.

[31] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 448–457, 2008.

[32] W.-S. Lee, Y.-W. Roh, D.-J. Kim, J.-H. Kim, and K.-S. Hong, "Speech emotion recognition using spectral entropy," in *Intelligent Robotics and Applications*. Springer, 2008, pp. 45–54.

[33] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *Selected Areas in Communications, IEEE Journal on*, vol. 6, no. 2, pp. 314–323, 1988.

[34] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2007, vol. 22.

[35] M. E. Sánchez-Gutiérrez, E. M. Albornoz, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, "Deep learning for emotional speech recognition," in *Pattern Recognition*. Springer, 2014, pp. 311–320.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[37] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.