Embedding Stacked Bottleneck Vocal Features in a LSTM Architecture for Automatic Pain Level Classification during Emergency Triage

Fu-Sheng Tsai
Department of Electrical Engineering
National Tsing Hua University, Taiwan
Email: eason820316@gmail.com

Yi-Ming Weng
Department of Emergency Medicine
Chang Gung Memorial Hospital, Taiwan
Email: wengym33@gmail.com

Chip-Jin Ng
Department of Emergency Medicine
Chang Gung Memorial Hospital, Taiwan
Email: ngowl9@adm.cgmh.org.tw

Chi-Chun Lee
Department of Electrical Engineering
National Tsing Hua University, Taiwan
Email: cclee@ee.nthu.edu.tw

***Abstract*—In order to effectively allocate healthcare resource, a proper triage classification system plays an important role in assessing the severity of on-boarding patients at the emergency department. One of the major items in the current triage system is to assess the level of pain intensity, which relies solely on patients self-report numerical-rating scale (NRS) at the moment. The nature of self-report on pain level poses a challenge in maintaining the validity and consistency of the triage classification outcome. While there has been algorithms developed to automatically detect pain from expressive behaviors, most of them concentrate only on facial or body gestural expressions within the context of physical exercises. In this work, we propose to utilize stacked bottleneck acoustic representations in a long-short term memory neural networks (LSTMs) architecture as features for pain severity classification in a database consists of patients during real triage sessions. Our proposed framework achieves accuracy of 72.3% and 54.2% in binary and three-class pain intensity classification tasks. Our results further demonstrate that the severity of pain can largely be captured in the patients prosodic characteristics.**

1. Introduction

Research in developing computational behavior analytics from measurable signals, e.g., audio-video and/or physiological data recordings, offer a new paradigm of quantitative decision-making for the domain experts [1]. Development of these behavior analytics are often grounded in their desired applications by providing consistent and objective data-driven measures of humans internal attributes. For example, computational advancements have already been observed in various applications within the medical domain: detection of depression [2], [3], assessment of Parkinsons disease [4], [5], modeling of therapists empathy in motivational interviews [6], [7], analysis of autism spectrum disorder [8], [9]. In this work, we carried out a collaborative behavioral signal processing (BSP) research effort with medical professionals toward automatically classifying pain level intensity of an on-boarding emergency patients by modeling their vocal characteristics during triage.

The Taiwan Triage and Acuity Scale (TTAS) [10] is jointly developed by the Taiwan Society of Emergency Medicine and the Critical Care Society, which modifies the Canadian Triage and Acuity Scale (CTAS) [11] to tailor toward Taiwan's particular medical situations. It is officially announced in 2010 by the Ministry of Health and Welfare to be the triage system of Taiwan. The pain level is one of six major regulators in the TTAS. While there has been a number of assessment tools developed for measuring pain in the medical domain, NRS, i.e., a 10-point self-report numerical-rating pain scale, remains to be the gold standard used in clinical practices [12], [13]. However, triage nurses have noticed challenges in the practical implementation of NRS especially for elderly people, foreigners, or patients with low education level; this self-report rating additionally suffers from various unwanted idiosyncratic factors, e.g., age and body part dependency and inconsistent comprehension of the pain scale. These issues often cause a deviation in the validity of the emergency triage classification. As a result, A joint collaborative work is initiated in order to objectify measures of pain intensity during triage by modeling patients multimodal behavior signals [14].

Most of the past engineering works in automatic recognition of pain have concentrated mainly by monitoring facial expressions or body gestures. For example, Littlewort et al. showed that by capturing a subject's 20 facial action units from 26 video recordings, they were able to classify between real and fake pain [15]. Guanming et al. proposed to extract weighted local binary attern (LBP) as features to recognize four different states of neonatal (calm, crying, moderate pain, and severe pain) [16]. Kaltwang et al. utilized the active appearance model (AAM) to quantify the subject's facial expressions in the UNBC-McMaster Shoulder-Pain dataset of 25 subjects in order to automatically classify pain versus no pain [17]. Aside from facial expressions, researchers have also modeled body gestures and motion descriptors in addition to facial expressions for pain detection [18], [19]. The heavy focus on facial expressions and body movements is partially due to the fact these works have concentrated in situations where the subjects are

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

being induced for pain experiences by performing physical exercises.

Our goal is to model the pain intensity during emergency triage where there is a natural spoken interaction between a medical professional and a patient, not only the facial expressions and the body gestures are available, but also the verbal behaviors can be captured and analyzed. In fact, the previous result indicated that vocal characteristics possess substantial information about the pain-intensity level [14]. In this work, given the recent success of utilizing long-short term memory neural networks (LSTMs) in obtaining the state-of-art recognition accuracy across applications, we develop a novel framework in leveraging stacked bottleneck acoustic features using LSTMs for automatic pain classification. Deep bottleneck features (DBFs) are generated based on using deep neural networks with a structure in learning a relatively narrower bottleneck hidden layer in order for form a low-dimensional compressed representation of the original input sequence features. Past works have demonstrated the successful usages of DBFs in a variety of learning tasks. For example, Song et al. proposed an improved i-vector representation based on DBFs for language identification [20]. Haag et al. proposed to use stacked bottleneck features and Bi-directional LSTMs for expressive head motion synthesis in spoken dialogs between actors; embedding stacked bottleneck features architecture in modeling context and expressive variability results in a significant improvement over conventional feed-forward deep neural networks [21].

In this work, we derive stacked bottleneck acoustic features by first pre-training an unsupervised sequence-to-sequence LSTM autoencoder on a background Chinese corpus and further fine-tuning it on the emergency triage database. The fine-tuned output hidden layers are fed into support vector machine classifier in order to perform the final pain level classification. Our proposed framework achieves a 72.3% accuracy in classifying between the extremes (severe versus mild) pain level and 54.2% accuracy in performing a three-class (severe, moderate, and mild) pain level recognition using only prosodic features. We also observe that the framework trained with prosodic low-level descriptors (LLDs) outperform spectral-based (MFCCs) LLDs, which implicates that the level of pain experienced may be related more to the prosodic structures and voice qualities of vocal expressions. The rest of the paper is organized as follows: section 2 describes about data collection, deep bottleneck vocal features, and pain classification, section 3 includes experimental setups and results, and section 4 concludes with future work.

2. Research Methodology

2.1. Triage Pain-Level Multimodal Database

The database included audio-video recordings, physiological (heart rate, systolic and diastolic blood pressure) vital sign data, and other clinically-related outcomes of onboarding emergency patients collected at the Department of Emergency at Chang Gung Memorial Hospital[1]. We excluded pediatric, trauma patients, and referral patients or patients with prior treatment before arrival, and we restricted our inclusion criterion as patients with symptoms of chest, abdominal, lower-back, limbic pain, and headaches. There were two sessions recorded for each patient, i.e., at initial triage and follow-up, where the follow-up session occurred approximately 1 hour after the treatment, if any, was given to the patient. These sessions essentially involved nurses asking the patient for the location of the body pain, the NRS scale of pain intensity (0-10, where 10 means the worst pain ever), and a brief description on the type of pain felt (for example, cramps or aches); it usually lasted around 30 seconds for each session. The audio-video data was recorded using a Sony HDR handy cam on a tripod in a designated assessment room, attempting to capture the patients vocal and facial expressions.

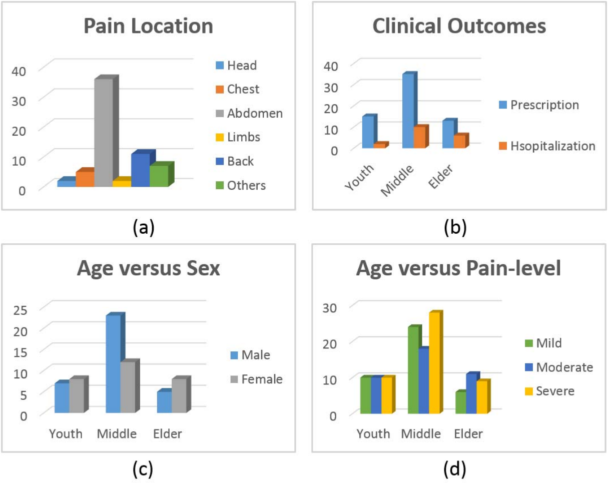

Figure 1. *Demographics and the NRS pain level distribution of patients used in this work. (a) Pain location distribution. (b) Clinical outcomes of analgesic prescription and hospital disposition (c) The distribution of age and gender. (d) The distribution of NRS pain level versus age*

Since the reliability of the NRS is crucial in the development and evaluation of our automated classification framework, we used a subset of the entire database in this pilot work. There are a total of 63 patients reporting a decrease in the NRS pain level between the initial and the follow-up triage sessions after being clinically-intervened with an analgesic prescription. This set of 126 samples can be seen as the set of samples whose self-reported pain levels are in accordance with the intended clinical validity in the development of NRS in assessing pain, i.e., a patient should report verbally a relieve in the pain symptoms after being medically treated. Hence, we use this particular set of the samples as the dataset of interest for this pilot algorithmic development work. Further, we categorize the NRS into three commonly-used pain levels, i.e., severe vs. moderate vs. mild. Severe pain corresponds to the NRS score ranging between 7-10, moderate is 4-6, and mild is 0-3. We set up two different recognition experiments in this work: 1) binary classification between the extreme pain levels, i.e., severe vs. mild pain, and 2) three-class pain-level classification.

[1] IRB#:CM104-3625B

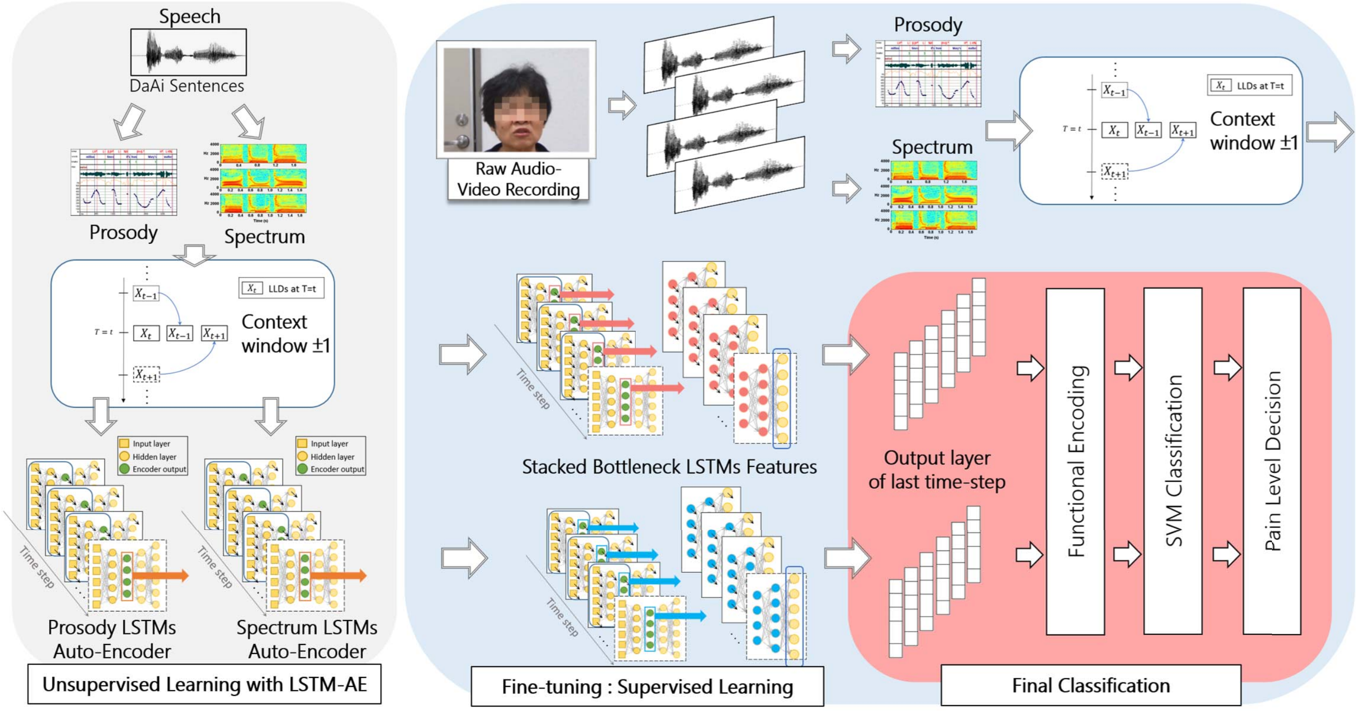

Figure 2. *It shows the complete architecture of our stacked bottleneck vocal feature computation used for automatic pain classification: unsupervised learning with LSTM autoencoder to extract the bottleneck layer, fine-tuning the bottleneck layer on the patient's acoustic data with supervised learning, and performing recognition with the fine-tuned outputted layer after functional encoding using support vector classification.*

2.2. Bottleneck Vocal Features with LSTM

Figure 2 shows the complete architecture of our stacked bottleneck vocal feature architecture used for automatic pain classification: pre-training LSTM autoencoder to extract the bottleneck layer, fine-tuning the bottleneck layer on the triage audio data with supervised learning, and performing recognition using the fine-tuned outputted layer. In this work, we choose to use LSTM due to its state-of-art capabilities in modeling long-term temporal dependencies and help avoid the problem of vanishing gradient as compared to recurrent neural network (RNN) [22]. Past work has also demonstrated the effectiveness of deriving bottleneck features using LSTM [23].

A typical LSTM is a time series model consists of forget gate f_t, input gate i_t, hidden/control state h_t, update gate z_t, reset gate r_t, and memory cell state C_t at every time step t:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

where $\tilde{C_t}$ can be seen as a candidate memory cell state used to generate the memory cell state C_t with the information gathered from the previous cell state C_{t-1}.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C_t}$$

The output will be based on cell state and shown as follows:

$$o_t = \sigma(Wo \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

Cho et al. [24] introduced Gate Recurrent Unit, or GRU, a more dramatic variation on LSTM. It combines the input and forget gate into signal gate which is defined as $z_t = \sigma(Wz \cdot [h_{t-1}, x_t])$. The resulting model is widely used and simpler than standard models. Finally, the output hidden state is defined as follows:

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h_t}$$

$$r_t = \sigma(Wr \cdot [h_{t-1}, x_t])$$

$$\tilde{h_t} = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

Thus, given an input sequence $X(x_1, ..., x_n)$, the above LSTM architecture calculates and outputs the hidden states $(h_1, ..., h_n)$. Finally, it defines a distribution over the output sequence $Y(y_1, ..., y_n)$ give the input sequence $X(x_1, ..., x_n)$, $p(Y \mid X)$, as below:

$$p(y_1, ..., y_m \mid x_1, ..., x_n) = \prod_{t=1}^{m} p(y_t \mid h_{n+t-1}, y_{t-1})$$

where $p(y_t \mid h_{n+t})$ is given by the softmax function.

2.2.1. Vocal Acoustic Low-level Descriptors (LLDs).

We extract two different types of input vocal descriptors generated in Praat [25] at a framerate of 10ms in this work:

- *Prosodic* features: pitch, intensity and harmonic-to-noise ratio (HNR), their delta and delta-delta
- *Spectral* features: 13 MFCCs and their delta and delta-detla

Then for every frame t, we perform context expansion by using a window of $(t + 1, t - 1)$ resulting in a total of

27 dimensional low-level descriptors as input sequence X for prosodic feature-based LSTM and 117 dimensional low-level descriptors for spectral feature-based LSTM.

2.2.2. Unsupervised Pre-training LSTM Autoencoder. We train an unsupervised LSTM sequence-to-sequence autoencoder on a background Chinese corpus, i.e., the DaAi database. The DaAi database is a large collection of Mandarin Chinese TV talk shows roughly includes 500 hours of speech data. In this data, we select only a subset of the DaAi containing 17,084 sentences totaling around 12 hours of audio samples to train the LSTM sequence-to-sequence autoencoder.

In particular, we pre-train a prosody-AE and spectral-AE separately. Both autoencoders include 5 total hidden layers of encoder-decoder; the prosody-AE has encoder-decoder hidden layers with the number of nodes set at 27, 15, 8, 15, 27, and we set the number of nodes at 117, 80, 32, 80, 117 for the spectral-AE. The loss function used here is the mean squared error, and the epoch used is 20.

2.2.3. Fine-tuning Stacked Bottleneck Layer. We then encode each of the spoken sentences of each patient in the triage database to the pre-trained bottleneck middle layer (8 nodes for prosody-AE and 32 nodes for spectral-AE). Then, we perform fine-tuning on these bottleneck layers by adding two additional layers to their corresponding LSTM (prosody-AE: 8, 32, 62, and spectral-AE: 32, 64, 128) with binary and sparse categorical cross entropy loss function for binary and three-class recognition tasks, respectively. The epoch is set at 20. The output layer of the last time step at the sigmoid and softmax layer for the binary and three-class learning respectively is extracted as patients acoustic representation at the sentence-level. These sentence-level features are then fed into another layer of encoding in order to perform the final pain classification for each patient for a triage session.

2.3. Pain Level Classification

Since every triage is of different length resulting in a varying number of sequences outputted from section 2.2.3. We compute 15 different statistical functionals on the hidden output to generate the final feature vector of each patient for every triage session. The list of functionals includes maximum, minimum, mean, median, standard deviation, 1st percentile, 99th percentile, 99th − 1st percentile, skewness, kurtosis, minimum position, maximum position, lower quartile, upper quartile and interquartile range. The selected classifier for training and recognition is linear-kernel support vector machine.

3. Experimental Setups and Results

We report recognition results on 1) binary classification between the extreme pain levels and 2) three-class pain-level classification in this section. Accuracies are measured in unweighted average recall (UAR) with the evaluation scheme done via leave-one-patient-out cross-validation. Except for the pre-training, the rest of the learning steps are carried out in the training set only.

3.1. Baseline Systems

We compare our method to two different baseline methods. The prosody-based baseline is done by first computing 6 prosodic features per frame (pitch, intensity and their associated delta and delta-delta) on the patient's speaking portion of each triage session. Further, we perform session-level encoding of these frame-level features using a representational learning technique, the Gaussian Mixture Model based Fisher Vector (GMM-FV) [26] - a method that has recently been shown to perform competitively in tasks of paralinguistic recognition using speech acoustics [27]. A brief description is below:

For a frame-level data sequence X, we can define a scoring function:

$$G_\lambda^X = \triangledown_\lambda log u_\lambda(X)$$

where $u_\lambda(X)$ denotes the likelihood of X given the probability distribution function (PDF). We use GMM as our PDF. λ represents the parameters of GMM, $\lambda = w_k, u_k, \sum_k, k = 1, ..., K$. G_λ^X is the direction where λ has to move to provide a better fit between u_λ and X. Fisher vector encoding is derived by computing the following first and second order statistics:

$$g_{u_k}^X = \frac{1}{T\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{x_t - u_k}{\sigma_k} \right)$$

$$g_{\sigma_k}^X = \frac{1}{T\sqrt{2w_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{(x_t - u_k)^2}{\sigma_k^2} - 1 \right)$$

$\gamma_t(k)$ is defined as

$$\gamma_t(k) = \frac{w_k u_k(x_t)}{\sum_{j=1}^{K} w_j u_j(x_t)}$$

The concatenation of $[g_{u_k}^X g_{\sigma_k}^X]$ is the session-level feature encoding that is used as the input to support vector classifier. We also report accuracy on spectral-based baseline, which computes 39 MFCCs (13 coefficients with delta and delta-delta) as the LLDs instead.

3.2. Results and Discussions

Table 1 summarizes our recognition results on the two NRS pain level recognition tasks, i.e., the binary classification between the extremes and ternary classification between the three pain levels. Prosody and MFCC indicate baseline systems (section 3.1) using prosody and MFCCs features respectively; DBAE-Prosody and DBAE-MFCC denote our proposed stacked bottleneck LSTM recognition framework. The best accuracy obtained in binary classification is the MFCC baseline (73.3%) where the best DBAE approach is the DBAE-Prosody (72.3%). In the more complex three-class recognition problems, our proposed DBAE-Prosody approach performs the best (54.2%) out of these four model. One thing to note that since spectral features can be highly sensitive to environmental factors, it is promising to see that DBAE-Prosody, which relies only on a few number of low level prosodic features, not only consistently outperforms DBAE-MFCC but also obtains stable and reliable accuracies across both binary and ternary recognition tasks. It may also

TABLE 1. *It summarized the Unweighted Averaged Recall (UAR) obtained in our pain-level recognition experiment. 2-Class indicated the binary classification task between the extreme pain levels. 3-Class indicated the ternary classification between severe, moderate, and mild pain levels. The number in bold indicated the best accuracy achieved within proposed framework. DBAE denoted Deep Bottleneck Feature Auto-Encoder Architecture*

2-Class	**Prosody**	**MFCC**	**DBAE- Prosody**	**DBAE- MFCC**	**Prosody+ MFCC**	**DBAE-Prosody+ DBAE-MFCC**
Mild	77.5	70.0	85.0	62.5	67.5	67.5
Severe	68.1	76.6	59.6	74.5	76.5	80.9
UAR	72.8	73.3	72.3	68.5	72.0	**74.2**
3-Class						
Mild	60.0	50.0	60.0	52.5	50.0	45.0
Moderate	41.0	46.2	28.2	41.0	46.2	38.5
Severe	51.0	61.7	74.5	57.4	55.3	68.1
UAR	50.7	52.6	**54.2**	50.3	50.5	50.5

points to the fact that a high degree of pain-related vocal characteristics are reflected in the prosodic characteristics of the patients.

We further perform recognition using a combination of the two set of features (Prosody and MFCC) with late fusion technique, i.e., by fusing the decision scores outputted from the prosodic and spectral systems separately using linear support vector classifier. The best accuracies obtained in the binary classification is by fusing DBAE-Prosody with DBAE-MFCC (74.2%), and the fusion of these two features actually degrade the UAR to 50.5%. The use of MFCCs need to be taken with caution since they are sensitive to the recording conditions which may reflect differences in the initial triage versus the follow-up triage instead of the actual pain-level reported. however, additional detailed studies are required to understand the effect. In summary, we observe that our proposed novel computational framework of deep bottleneck feature using LSTM auto-encoder architecture, especially when trained on prosodic features, can obtain promising recognition accuracies in classifying the levels of NRS pain scale in a real emergency triage session.

3.2.1. Additional Analysis. We additionally present different accuracies obtained in the two recognition tasks by altering the number of additional fine-tuning layers added to the extracted bottleneck layer in DBAE-Prosody and DBAE-MFCCs in Figure 3. The legend in Figure 3 indicates the number of nodes in the outputted layers of final LSTM. For DBAE-Prosody, the best accuracies obtained plateaus when adding two additional layers in both 2-class and 3-class tasks, and the same trend holds for DBAE-MFCC in the 2-class recognition task.

4. Conclusions

Due to the inconsistency and subjective nature of the currently implemented NRS pain scale, our aim is to develop an objective method in measuring pain-level intensity of patients during emergency triage. In this work, we propose a novel computational framework in embedding bottleneck vocal features in a LSTM architecture to automatically recognize pain-level intensity for emergency room patients during triage. The encoded bottleneck vocal features are trained through unsupervised sequence-to-sequence autoencoder using a background corpus, which are then fine-tuned on the target triage database. We demonstrate that this novel approach especially DBAE-Prosody, which learns only based on a few prosodic features, is capable of obtaining promising accuracies in both 2-class and 3-class recognition task. To the best of our knowledge, this is one of the first works that have used voice characteristics in detecting pain and have contextualized such an effort in a database collected out of real patients.

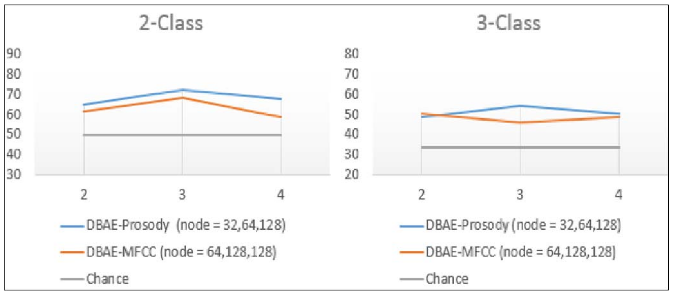

Figure 3. *It shows the recognition accuracies obtained with a varying number of additional hidden layers added in the fine-tuning of bottleneck features. X-axis is the number of hidden layer, and Y-axis is UAR.*

While the initial result is quite promising, there are multiple future directions to pursue. One of the immediate directions is to expand the scale of the current databases, i.e., both in the collection of the actual triage sessions (our aim is to collect at least 500 unique patients data) and also in the utilization of the DaAi background corpus, to enrich the behavior variabilities observed and modeled in our current framework. Secondly, we would develop a joint multimodal framework to include facial expressions and body gestures into our classification system. On the analysis part, we will investigate further the robustness of prosodic features in characterizing the levels of pain intensity and also conduct research into understanding not only the expressive behaviosr but also the internal physiology (blood pressure and heart rate) in relation to the pain levels. Lastly, pain has recently been conceptualized as a homeostatic emotion [28], which points potentially toward the use affective computing technology as additional sources of information to characterize this internal attribute of human. The overarching goal is to finally develop an objective and quantifiable clinically-valid measure of pain not only to replicate the current pain-level assessment instruments but to provide supplementary information that is beyond the established protocols to enhance the effectiveness of emergency triage classification.

Acknowledgments

Thanks to Ministry of Science and Technology (103-2218-E-007-012-MY3) and Chang Gung Memorial Hospital (CMRPG3E1791) for funding.

References

[1] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[2] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–7.

[3] Z. Liu, B. Hu, L. Yan, T. Wang, F. Liu, X. Li, and H. Kang, "Detection of depression in speech," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 743–747.

[4] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 181–190, 2014.

[5] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully automated assessment of the severity of parkinson's disease from speech," *Computer speech & language*, vol. 29, no. 1, pp. 172–185, 2015.

[6] J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. S. Narayanan, "Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms." in *INTERSPEECH*, 2015, pp. 1947–1951.

[7] B. Xiao, D. Can, P. G. Georgiou, D. Atkins, and S. S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–4.

[8] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Computer speech & language*, vol. 29, no. 1, pp. 132–144, 2015.

[9] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1162–1177, 2014.

[10] C.-J. Ng, Z.-S. Yen, J. C.-H. Tsai, L. C. Chen, S. J. Lin, Y. Y. Sang, J.-C. Chen, T. national working group *et al.*, "Validation of the taiwan triage and acuity scale: a new computerised five-level triage system," *Emergency Medicine Journal*, vol. 28, no. 12, pp. 1026–1031, 2011.

[11] M. J. Bullard, B. Unger, J. Spence, and E. Grafstein, "Revisions to the canadian emergency department triage and acuity scale (ctas) adult guidelines," *Cjem*, vol. 10, no. 02, pp. 136–142, 2008.

[12] K. Eriksson, L. Wikström, K. Årestedt, B. Fridlund, and A. Broström, "Numeric rating scale: patients' perceptions of its use in postoperative pain assessments," *Applied nursing research*, vol. 27, no. 1, pp. 41–46, 2014.

[13] E. Castarlenas, E. Sánchez-Rodríguez, R. de la Vega, R. Roset, and J. Miró, "Agreement between verbal and electronic versions of the numerical rating scale (nrs-11) when used to assess pain intensity in adolescents," *The Clinical journal of pain*, vol. 31, no. 3, pp. 229–234, 2015.

[14] F.-S. Tsai, Y.-L. Hsu, W.-C. Chen, Y.-M. Weng, C.-J. Ng, and C.-C. Lee, "Toward development and evaluation of pain level-rating scale for emergency triage based on vocal characteristics and facial expressions," *Interspeech 2016*, pp. 92–96, 2016.

[15] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1797–1803, 2009.

[16] G. Lu, C. Yang, M. Chen, and X. Li, "Sparse representation based facial expression classification for pain assessment in neonates," in *Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on*. IEEE, 2016, pp. 1615–1619.

[17] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," *Advances in visual computing*, pp. 368–377, 2012.

[18] M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh *et al.*, "The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 435–451, 2016.

[19] T. A. Olugbade, M. Aung, N. Bianchi-Berthouze, N. Marquardt, and A. C. Williams, "Bi-modal detection of painful reaching for chronic pain rehabilitation systems," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 455–458.

[20] Y. Song, B. Jiang, Y. Bao, S. Wei, and L.-R. Dai, "I-vector representation based on bottleneck features for language identification," *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, 2013.

[21] K. Haag and H. Shimodaira, "Bidirectional lstm networks employing stacked bottleneck features for expressive speech-driven head motion synthesis," in *International Conference on Intelligent Virtual Agents*. Springer, 2016, pp. 198–207.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2304–2308.

[24] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[25] P. Boersma and D. Weenink, "Praat-a system for doing phonetics by computer [computer software]," *The Netherlands: Institute of Phonetic Sciences, University of Amsterdam*, 2003.

[26] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," *Computer Vision–ECCV 2010*, pp. 143–156, 2010.

[27] H. Kaya, A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis." in *INTERSPEECH*, 2015, pp. 909–913.

[28] A. Craig, "A new view of pain as a homeostatic emotion," *Trends in neurosciences*, vol. 26, no. 6, pp. 303–307, 2003.