Interspeech 2018
2-6 September 2018, Hyderabad

Learning Conditional Acoustic Latent Representation with Gender and Age Attributes for Automatic Pain Level Recognition

Jeng-Lin Li[1,5], *Yi-Ming Weng*[2,3,4], *Chip-Jin Ng*[2], *Chi-Chun Lee*[1,5]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Department of Emergency Medicine, Chang Gung Memorial Hospital, Taiwan
[3]Department of Emergency Medicine, Tao-Yuan General Hospital, Taiwan
[4]Faculty of Medicine, National Yang-Ming University, Taiwan
[5]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

[1]cllee@gapp.nthu.edu.tw, [1]cclee@ee.nthu.edu.tw

Abstract

Pain is an unpleasant internal sensation caused by bodily damages or physical illnesses with varied expressions conditioned on personal attributes. In this work, we propose an age-gender embedded latent acoustic representation learned using conditional maximum mean discrepancy variational autoencoder (MMD-CVAE). The learned MMD-CVAE embeds personal attributes information directly in the latent space. Our method achieves a 70.7% in extreme set classification (severe versus mild) and 47.7% in three-class recognition (severe, moderate, and mild) by using these MMD-CVAE encoded features on a large-scale real patients pain database. Our method improves a relative of 11.34% and 17.51% compared to using acoustic representation without age-gender conditioning in the extreme set and the three-class recognition respectively. Further analyses reveal under severe pain, females have higher maximum of jitter and lower harmonic energy ratio between F0, H1 and H2 compared to males, and the minimum value of jitter and shimmer are higher in the elderly compared to the non-elder group.
Index Terms: pain, acoustic representation, age and gender, conditional variational autoencoder (CVAE)

1. Introduction

Pain is a subjective internal sensation, and its intensity is often related to past personal experiences of painful episodes. These episodes are associated mostly with bodily damages or physical illnesses. Research has shown that the self-reported pain levels and the biological responses to pain induced stimuli are dependent on an individual's personal attributes (age and gender) [1]. For example, fenale tend to report a higher-level of pain, i.e., a lower pain tolerance [2, 3]; some research ascribes this to the societal stereotypes on the expected pain endurance of being feminism or masculine [4, 5, 6]. In terms of age, elder people have also been shown to have lower pain tolerance threshold [1, 7, 8]. This age-dependent phenomenon is theorized to be caused by multi-dimensional factors (sensory, affect and cognition) resulting in a modified psychological strategy for the elderly to appraise pain [9]. Empirical evidences of gender-dependent biological responses to pain stimuli have also been reported: pupil dilations occurs more in females than males when experiencing a high intensity pain [10]; the differences between gender is not only observed in their reported pain levels but also seen in the brain-related neural responses [11, 12].

While being a complex internal sensation, much effort has been devoted in developing strategies for objectively assessing pain levels due to its significant implications of survival and resource management in clinical applications [13]. For example, pain level is one of the major factors used in the triage and acuity scale for emergency department to screen life-threatening patients [14, 15]. Assessing pain also helps evaluate the effect of analgesia on postoperative, endodontic and multiple treatments [16, 17] and is essential in improving quality of healthcare [18]. Currently, the clinical gold standard in pain assessment is done via individual's self-reported numerical scale. This method is known to be unreliable for elderly, cognitively impaired, and young children [19]. The use of self-report further hinders the large scale medical applications requiring continuous pain-level monitoring.

Research has indicated that observational based measures are essential in achieving consistency in assessing pain [20]. In fact, several engineering effort has computed pain from measurable data. Most of these works focus on modeling facial expressions; for example, Rodriguez et al. uses Long Short-Term Memory Networks on images of face to estimate pain intensity [21], Zhang et al. proposes binary edge features to model 3D facial expression for pain expression [22], and Egede et al. combines deep learning features and hand-crafted features with a fusion scheme to predict pain intensity [23]. Only recently, Tsai et al. has initiated an investigation in performing pain level estimation from acoustic cues in triage settings [24, 25].

We propose to learn acoustic latent representations embedded with attributes of gender and age using conditional variational autoencoder optimized with criterion of maximum-mean discrepancy (MMD-CVAE) for automatic pain level recognition. Conventionally, the issue of personal attribute dependency is handled by training multiple independent models, e.g., gender or age specific, our proposed MMD-CVAE, however, directly embeds this information in the encoded acoustic space. We evaluate our method on a large-scale real patients database collected during triage sessions [24]. The use of our proposed MMD-CVAE encodes acoustic representation in training pain level classifier achieves 70.7% in extreme set classification (severe versus mild) and 47.7% in three-class recognition (severe, moderate, and mild). Deriving representation by learning with personal attributes condition help improve the pain level recognition for a relative gain of 11.34% and 17.51%. Further analyses demonstrate that females exhibit higher maximum of jitter and lower harmonic energy ratio between F0, H1 and H2 compared to males patients while the minimum value of jitter and shimmer are higher in the elderly than in the non-elder group.

The rest of the paper is organized as follows: Section 2 describes data collection and framework. Section 3 shows our experimental results. Section 4 concludes with future work.

10.21437/Interspeech.2018-1298

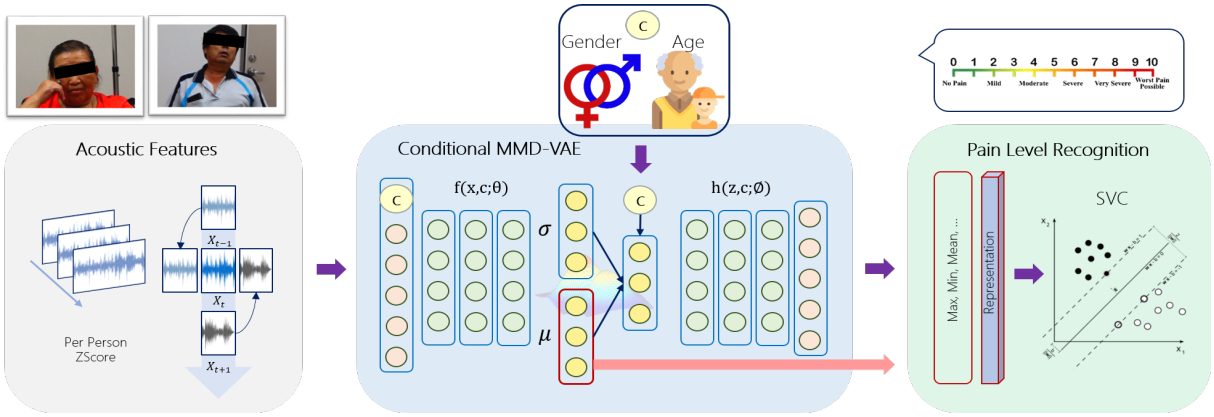

Figure 1: *This is our overall framework. A conditional variational autoencoder architecture with maximum-mean-discrepancy criterion is used to learn acoustic representation for automatic pain classification. The acoustic representation is derived by encoding frame-level acoustic low-level descriptors using the learned $f(x, c; \theta)$. The pain-level classifier is then trained on session-level features computed by using statistical functions on these conditional latent representations*

2. Research Methodology

2.1. The Triage Pain-Level Audio-Video Database

The pain-level database consists of data from real on-boarding patients collected during triage sessions at the Chang Gung Memorial Hospital Emergency Department [1]. These sessions involves triage nurse engaging in spoken interactions with the patient in order to record the patient's NRS pain scale, i.e., a 10-point self-report numerical-rating pain scale (0-10, where 10 means the worst pain ever), location of pain, and a brief description of the pain felt. Each session lasts approximately 30 seconds. During the session, we collect both audio-video recordings (using a Sony HD camcorder) and other relevant information (physiological vital sign, personal information, and clinically-relevant outcomes). Each patient undergoes the session twice - one at pre-intervention and one at post-treatment.

After excluding low quality samples (either low audio-video quality or missing personal attributions or clinical outcomes), there are 141 unique patients with a total of 335 unique sessions. We utilize this entire dataset in our work as compared to the previous works on the same database, where only subset of this database were included [24, 25]. Figure 2 shows a breakdown of pain levels for each personal attribute (age and gender) in this entire database. The pain intensity is categorized into three levels based on the reported NRS score, i.e, 0-3 mild, 4-6 moderate, and 7-10 severe; age is also categorized into three level, i.e., youth (0-40), middle-age (41-64), and elder (over 65). There are 201 sessions of male patients and 134 of female patients. Male's average pain-intensity is 4.821 and female is 5.190, and the average pain level is 4.648, 5.057 and 5.167 for youth, middle-age and elder, respectively.

2.2. Vocal Representation using MMD-CVAE

In this work, we learn a conditional generative neural network using VAE with MMD criterion at the frame level. The conditioning is given with age and gender attributes. The learned MMD-CVAE encoder network can then be used to compute acoustic latent representation. Both components of acoustic LLDs and MMD-CVAE are described as following.

2.2.1. Vocal Acoustic Low-level Descriptors (LLDs)

As eGeMAPS acoustic low-level descriptor set serves as one of an effective set of acoustic parameters across different affect-based recognition tasks [26], we extract it for the pain-level recognition task. The parameter setting including is the same as [26] except that we discard MFCCs (the first 4 coefficients) due to their high sensitivity to noisy conditions. The features are further z-normalized per speaker. Then, we apply context window expansion to extend every frame t with contextual information from two neighboring frames, $(t+1, t-1)$. This results in a total of 57 dimensional low level descriptors.

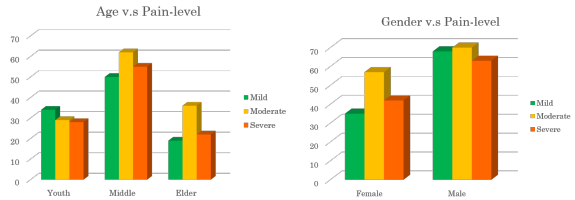

Figure 2: *A summary of pain level distribution as a function of personal attributes (age and gender) in the triage pain-level audio-video database.*

2.2.2. MMD-CVAE

We encode acoustic LLDs into latent representations using conditional VAE. Conventional VAE relying on Kullback-Leibler Divergence (KLD) by optimizing the evidence lower bound (ELBO), but it suffers from a serious issue that the learned latent representation may not be informative [27, 28]. To embed fine-grained information such as gender and age attributes in the acoustic latent space, we utilized the Maximum-Mean Discrepancy Variational Autoencoder (MMD-VAE) that has been shown to alleviate the problem of uninformative latent representation [29]. The gender and age attributes are embedded as a probabilistic conditioning in learning the MMD-VAE network.

An input x can be encoded with a VAE to derive a latent code z using the learned encoding network f. f models $p_\theta(x|z, c)$ parameterized by θ given personal attribute c. Since the true posterior distribution $p(z|x, c)$ is unknown, we can use a variational approach by defining $q_\phi(z|x, c)$. Then, we can learn the network weights by minimizing $D_{KL}(q_\phi(z|x, c)||p(z|x, c))$ which is simplified as an objective maximizing the following variational evidence lower bound ($\mathcal{L}_{ELBO}$):

$$\mathcal{L}_{ELBO} = \mathcal{L}_{KL} + \mathcal{L}_{Rec} \leqslant p(x|c) \quad (1)$$

$$\mathcal{L}_{KL} = -D_{KL}(q_\phi(z|x, c)||p(z|c)) \quad (2)$$

[1] IRB#:CM104-3625B

Table 1: *A summary of pain level recognition results. We include Func, VAE and CVAE approaches as our comparison methods. Func denotes LLDs directly encoded by functionals. The G and A indicate gender and age attributes respectively.*

2-Class	Func	VAE	Func-G	Func-A	VAE-G	VAE-A	VAE-(G+A)	CVAE-G	CVAE-A	CVAE-(G+A)
Mild	0.629	0.660	0.650	0.612	0.641	0.612	0.631	0.631	0.709	0.689
Severe	0.619	0.610	0.600	0.590	0.590	0.638	0.667	0.657	0.629	0.724
UAR	0.624	0.635	0.625	0.601	0.616	0.625	0.649	0.644	0.669	**0.707**
3-Class										
Mild	0.476	0.437	0.456	0.350	0.466	0.456	0.505	0.427	0.495	0.534
Moderate	0.409	0.354	0.291	0.496	0.370	0.433	0.323	0.370	0.480	0.449
Severe	0.352	0.381	0.410	0.467	0.381	0.305	0.390	0.410	0.419	0.438
UAR	0.383	0.391	0.386	0.437	0.406	0.398	0.406	0.402	0.465	**0.474**

$$\mathcal{L}_{REC} = \mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z,c)] \quad (3)$$

where the regularization term $\mathcal{L}_{KL}$ is encouraged to make $q_\phi(z|x,c)$ closer to the true prior $p(z|c)$. The use of KLD in this optimization approach can lead to the learned p_θ and q_ϕ to neglect latent code z, i.e., z and x are consequently independent. In this work, when learning the VAE, we replace $D_{KL}(q_\phi(z|x,c)||p(z|c))$ with distribution distance computed by Maximum-Mean Discrepancy (MMD):

$$D_{MMD}(q_\phi(z|x,c)||p(z|c)) = \mathbb{E}_{p(z|c),p(z'|c')}[k(z,z')]$$

$$-2\mathbb{E}_{q(z|x,c),p(z'|c')}[k(z,z')] + \mathbb{E}_{q(z|x,c),q(z'|x',c')}[k(z,z')]$$

where $k(z,z')$ is an positive definite kernel as $e^{-||z-z'||_2}$ and the $l2$ norm term is empirically divided by the dimension of z. $D_{MMD} = 0$ if and only if $p = q$.

In summary, we encode acoustic low level descriptors at frame level using f encoding network (probability distribution, $p_\theta(x|z,c)$) given patients personal attributions c (age and gender) to derive patients latent acoustic representation to be further used to train a classifier to perform pain level classification.

2.3. Pain Level Classification

Since each triage session lasts about 30 seconds, the latent frame-level features derived from section 2.2.2 are further encoded using functionals to obtain the high dimensional session level representation. We use 15 statistical functionals: 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. Finally, we perform pain level recognition using linear-kernel support vector machine with univariate feature selection.

3. Experimental Setups and Results

3.1. Experimental Setup

In this work, we present recognition results on two different tasks: 1) binary (2-class) classification between the extreme pain levels and 2) three-class (3-class) pain-level classification. The experiment is carried out using leave-one-speaker-out cross-validation with unweighted average recall (UAR) as evaluation metric. Univariate feature selection based on ANOVA test is also carried out.

Our MMD-VAE network includes three hidden layers with one latent layer. The network architectures used in this work are 55-40-40-35-40-40-55 and 50-50-40-35-40-50-50 for two class and three class tasks respectively We use leaky ReLU activation function in the hidden layers of encoder network and ReLU in decoder network. All the hidden layers are batch normalized. The batch size is specified as 50 and the learning rate is 0.001 with Adam optimizer and 10 epochs network optimization.

3.1.1. Comparison Models

The following is the list of comparison models:

- **Func**: Model trained on directly encoding the LLDs with statistical functionals (similar to previous work [24])
- **VAE**: Model trained on encoding LLDs to MMD-VAE latent space
- **Func-X**: Same as Func approach but train a separate model for each X-attribute to generate multiple attribute-dependent models
- **VAE-X**: Same as VAE approach but train a separate model for each X-attribute to generate multiple attribute-dependent models
- **CVAE-X**: Model trained on encoding LLDs to MMD-CVAE conditioning on X-attribute

where *X* can be either gender (G) attribute, age (A) attribute, or gender-age (G+A) attributes jointly. For *Func-X* and *VAE-X*, there are 2 gender groups (male and female) which result in a male-specific model and a female-specific model; likewise, there are 2 age groups resulting in a non-elderly specific model (< 65) and an elderly specific model (> 65). When considering age-gender attributes jointly, there are 4 groups, i.e., the elder male, the elder female, the non-elder male, and the non-elder female. *CVAE-X* in our proposed model where the age-gender information is directly embedded in the learning of latent representation eliminating the needs for separate attribute-specific model training. Herein, *X* is the original integer values normalized by dividing 100 for age and the binary values for gender.

3.2. Pain-Level Recognition Results

Table 1 summarizes our pain-level recognition results. Our proposed conditional acoustic latent representation with gender and age attributes, *CVAE-(G+A)*, achieves the best accuracy of 0.707 and 0.474 in 2-class and 3-class classification, i.e., a relative gain of 11.34% and 17.51% over the *VAE* acoustic representation without age-gender conditioning. A few notable observations can be made from our comparison experiments.

Firstly, encoding LLDs into *VAE*'s (even without conditioning) improves recognition rates compared to methods based on directly computing functionals, *Func*. This result demonstrates that a more robust modeling of acoustic representations can be obtained by using generative variational autoencoder as feature extractor. Secondly, an intuitive method in handling the variability of different personal attributes is by training separate models, e.g., male and female separate models, non-elderly and elderly separate models. Table 1 shows that

Table 2: *A table summarizing our statistical analysis. Features that are significant different between groups of personal attribute are listed below. F denotes female, M denotes male, E is the elderly and NE is the non-elder people. All p-values are less than 0.01.*

Gender	Age
F>M	E>NE
(slope500-1500)-max	(logRelF0-H1-A3)-max
(jitter)-max	(F1amp)-max
(F1BW)-max	(F2amp)-max
(F3freq)-max	(F3amp)-max
(HNR)-min	(F0)-min
(F1freq)-min	(jitter)-min
(F2freq)-min	(shimmer)-min
(F3freq)-min	(F1amp)-min
	(F2amp)-min
	(F3amp)-min
F<M	E<NE
(F0)-max	(spectralFlux)-min
(logRelF0-H1-H2)-max	(F0)-median
(logRelF0-H1-A3)-max	(shimmer)-median
(HNR)-median	(logRelF0-H1-A3)-median
(logRelF0-H1-H2)-median	(F1amp)-median
(F1BW)-median	(F2amp)-median
	(F3amp)-median

Func-G has a slightly higher accuracy than *Func* in both tasks; *Func-A* improves *Func* in the 3-class classification task. The same trend also occurs in using attribute-specific model using *VAE* approaches, in specifics, the accuracy increases primarily in 3-class classification when using *VAE-G* and *VAE-A*, and by joint attributes-specific modeling, *VAE-(G+A)*, it obtain improvement on both tasks.

Finally, utilizing joint attributes modeling (G+A) is more beneficial in obtaining an improved recognition rate than single-attribute embedding. This is evident in both the *VAE* and our proposed approach *CVAE*, where the best accuracy also come from considering both attributes simultaneously. In summary, by directly embed the age-gender attribute information as conditional probability in the learning of generative VAE network, we can obtain a single latent representation that possess improved discriminatory power in detect pain from vocal cues.

3.3. Statistical Analyses

In this section, we present an statistical analysis in understanding the differences of patient's vocal characteristics as a function of personal attribute (age and gender) in a fixed category of pain intensity (severe pain). We first compute five different functionals (maximum, minimum, mean, median and standard deviation) on the extracted low-level acoustic descriptors, and we perform one-sided Student's t-tests ($\alpha = 0.01$ level) between two gender or age groups (male vs. female, or non-elder vs. elderly) suffering severe pain. Table 2 summarizes the results and the direction of differences. Note that logRelF0-H1-H2 is an abbreviation of the ratio of harmonic energy difference between first F0, harmonic (H1) and second harmonic (H2), A3 stands for the third formant range; amplitude, frequency and bandwidth are abbreviated as amp, freq and BW.

There are 14 and 17 out of 95 features that show statistically-significant differences between gender (males versus females) and age (elderly versus non-elder patients) groups when they are all experiencing severe pain. In specifics, the male patients suffering severe pain express pain with smaller maximum jitter than female patients do while the maximum and median of harmonic energy ratio between F0, H1 and H2 has much lower values among females. Likewise, the minimum value of jitter and shimmer is significantly larger in the elderly group than the non-elder patients when reporting severe pain. These LLDs are z-normalized with respect to individual speaker mitigating individual biases, such as male and female average pitch height. Therefore, it is interesting to see that our analysis results indicate that there indeed exists differences in the acoustic expressions among groups with different age and gender range, and proper handling them with our propose technical approach is beneficial in advancing pain recognition framework from speech modality.

4. Conclusions and Future Works

The difference in the vocal expressions of pain intensity is known to be related to an individual's personal attribute (e.g., age and gender). Leveraging the availability of personal attributes for accurate pain detection naturally fits well in real world clinical practices. In this work, we propose to learn a MMD-CVAE network embedded with personal attributes directly in the latent layer. By encoding acoustic LLDs into the MMD-CVAE latent representation, we improve the vocal-based pain-level recognition by using the derived age-gender attribute embedded vocal latent representations. We evaluate our framework on a large scale real patients triage audio-vido databases. Our proposed framework achieves 70.7% and 47.4% in recognizing self-reported pain levels, which improves over baseline models without such personal attributes conditioning. Our experiments also demonstrate that joint modeling on both age and gender outperform single attribute embedding; furthermore, we observe several acoustic patterns differs between gender and age groups. To our best knowledge, this is one of the first work in incorporating personal attributes directly via a conditional-generative autoencoder network in order to learn a attribute-meaningful latent representation to improve pain detection from acoustic features.

In our future work, we plan to continue our research in three major directions: modeling power, multimodal integration, and comprehensive data collection with personal attributes. Firstly, while MMD-CVAE is an advanced model to properly learned latent representation as an autoencoder, pain as an internal hidden sensation, the complexity in modeling its manifested behaviors remain to be challenging. Technically, we will investigate approaches in imposing a more complex (less regularized) prior hypothesis, such as mixture of Gaussian [30], to further increase the modeling power of our vocal-based autoencoder network. Secondly, developing multimodal framework conditioned on individual personal attributes that integrates facial expressions, acoustic features and lexical information data will help provide a more robust and reliable assessment. Thirdly, personal attributes are not only restricted to gender and age attributes, we plan to include other relevant patients profiles (those could be related to clinically-relevant meta data) as part of the latent representation learning. We will continue to extend our interdisciplinary effort opportunities in deriving human behavior analytics [31], especially focusing on the health applications.

5. References

[1] L. D. Wandner, C. D. Scipio, A. T. Hirsh, C. A. Torres, and M. E. Robinson, "The perception of pain in others: how gender, race, and age influence pain expectations," *The Journal of Pain*, vol. 13, no. 3, pp. 220–227, 2012.

[2] A. M. Unruh, "Gender variations in clinical pain experience," *Pain*, vol. 65, no. 2-3, pp. 123–167, 1996.

[3] W. Gutiérrez Lombana and S. E. Gutiérrez Vidál, "Pain and gender differences. a clinical approach," *Revista Colombiana de Anestesiología*, vol. 40, no. 3, pp. 207–212, 2012.

[4] N. Solheim, S. Östlund, T. Gordh, and L. A. Rosseland, "Women report higher pain intensity at a lower level of inflammation after knee surgery compared with men," *Pain Reports*, vol. 2, no. 3, p. e595, 2017.

[5] C. Leboeuf-Yde, J. Nielsen, K. O. Kyvik, R. Fejer, and J. Hartvigsen, "Pain in the lumbar, thoracic or cervical regions: do age and gender matter? a population-based study of 34,902 danish twins 20–71 years of age," *BMC musculoskeletal disorders*, vol. 10, no. 1, p. 39, 2009.

[6] R. B. Fillingim, C. D. King, M. C. Ribeiro-Dasilva, B. Rahim-Williams, and J. L. Riley, "Sex, gender, and pain: a review of recent clinical and experimental findings," *The journal of pain*, vol. 10, no. 5, pp. 447–485, 2009.

[7] S. J. Gibson and R. D. Helme, "Age-related differences in pain perception and report," *Clinics in geriatric medicine*, vol. 17, no. 3, pp. 433–456, 2001.

[8] S. Lautenbacher, M. Kunz, P. Strate, J. Nielsen, and L. Arendt-Nielsen, "Age effects on pain thresholds, temporal summation and spatial summation of heat and pressure pain," *Pain*, vol. 115, no. 3, pp. 410–418, 2005.

[9] L. Petrini, S. T. Matthiesen, and L. Arendt-Nielsen, "The effect of age and gender on pressure pain thresholds and suprathreshold stimuli," *Perception*, vol. 44, no. 5, pp. 587–596, 2015.

[10] W. Ellermeier and W. Westphal, "Gender differences in pain ratings and pupil reactions to painful pressure stimuli," *Pain*, vol. 61, no. 3, pp. 435–439, 1995.

[11] S. M. Berman, B. D. Naliboff, B. Suyenobu, J. S. Labus, J. Stains, J. A. Bueller, K. Ruby, and E. A. Mayer, "Sex differences in regional brain response to aversive pelvic visceral stimuli," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 291, no. 2, pp. R268–R276, 2006.

[12] J.-K. Zubieta, R. F. Dannals, and J. J. Frost, "Gender and age influences on human brain mu-opioid receptor binding measured by pet," *American Journal of Psychiatry*, vol. 156, no. 6, pp. 842–848, 1999.

[13] H. Breivik, P. Borchgrevink, S. Allen, L. Rosseland, L. Romundstad, E. B. Hals, G. Kvarstein, and A. Stubhaug, "Assessment of pain," *British journal of anaesthesia*, vol. 101, no. 1, pp. 17–24, 2008.

[14] R. Beveridge, B. Clarke, L. Janes, N. Savage, J. Thompson, G. Dodd, M. Murray, C. N. Jordan, D. Warren, and A. Vadeboncoeur, "Implementation guidelines for the canadian emergency department triage & acuity scale (ctas)," *Canadian association of emergency physicians*, 1998.

[15] M. J. Bullard, B. Unger, J. Spence, E. Grafstein, C. N. W. Group *et al.*, "Revisions to the canadian emergency department triage and acuity scale (ctas) adult guidelines," *Canadian Journal of Emergency Medicine*, vol. 10, no. 2, pp. 136–142, 2008.

[16] F. E. C. Pereira, I. L. Mello, F. H. d. O. M. Pimenta, D. M. Costa, D. V. T. Wong, C. R. Fernandes, R. C. Lima Junior, and J. M. A. Gomes, "A clinical experimental model to evaluate analgesic effect of remote ischemic preconditioning in acute postoperative pain," *Pain research and treatment*, vol. 2016, 2016.

[17] S. Attar, W. R. Bowles, M. K. Baisden, J. S. Hodges, and S. B. McClanahan, "Evaluation of pretreatment analgesia and endodontic treatment for postoperative endodontic pain," *Journal of endodontics*, vol. 34, no. 6, pp. 652–655, 2008.

[18] N. Wells, C. Pasero, and M. McCaffery, "Improving the quality of care through pain assessment and management," 2008.

[19] D. Brown, "Pain assessment with cognitively impaired older people in the acute hospital setting," *Reviews in pain*, vol. 5, no. 3, pp. 18–22, 2011.

[20] T. Hadjistavropoulos and K. Craig, "A theoretical framework for understanding self-report and observational measures of pain: a communications model," *Behaviour research and therapy*, vol. 40, no. 5, pp. 551–570, 2002.

[21] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE transactions on cybernetics*, 2017.

[22] X. Zhang, L. Yin, and J. F. Cohn, "Three dimensional binary edge feature representation for pain expression analysis," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–7.

[23] J. Egede, M. Valstar, and B. Martinez, "Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 689–696.

[24] 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." in *INTERSPEECH*, 2016, pp. 92–96.

[25] F.-S. Tsai, Y.-M. Weng, C.-J. Ng, and C.-C. Lee, "Embedding stacked bottleneck vocal features in a lstm architecture for automatic pain level classification during emergency triage," *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 313–318, 2017.

[26] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[27] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," *arXiv preprint arXiv:1611.02731*, 2016.

[28] S. Yeung, A. Kannan, Y. Dauphin, and L. Fei-Fei, "Tackling over-pruning in variational autoencoders," *arXiv preprint arXiv:1706.03643*, 2017.

[29] S. Zhao, J. Song, and S. Ermon, "Infovae: Information maximizing variational autoencoders," *arXiv preprint arXiv:1706.02262*, 2017.

[30] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 1965–1972.

[31] 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.