A TRIPLET-LOSS EMBEDDED DEEP REGRESSOR NETWORK FOR ESTIMATING BLOOD PRESSURE CHANGES USING PROSODIC FEATURES

Hao-Chun Yang[1], *Fu-Sheng Tsai*[1], *Yi-Ming Weng*[2,3,4], *Chip-Jin Ng*[2], *Chi-Chun Lee*[1]

[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, Prehospital Care Division, Tao-Yuan General Hospital, Taiwan
[4]Faculty of Medicine, National Yang-Ming University, Taiwan

ABSTRACT

Studies have shown that measures of personal physiology, e.g., blood pressure (BP) variation and heart rate variability (HRV), is closely related to a subject's psychological states and are being used regularly to track patients' health conditions in medical settings. The conventional method of monitoring physiology requires wearing specialized sensors or utilizing medical instruments, which hinders the ability of scalable and just-in-time monitoring of patients. In this study, we propose a triplet-loss embedded deep regressor network to predict changes of BP using expressive prosodic features for on-boarding emergency room patients between *pre-* and *post-* triage sessions. The framework achieves correlations of 0.419 and 0.386 in predicting changes in SBP (systolic blood pressure) and DBP (diastolic blood pressure) respectively, which is 26.1% and 17.3% relative improvement compared to DNN-regressors without triplet-loss embedding. Further correlation analyses on the relationship between prosodic features and BP changes are presented.

Index Terms— behavioral signal processing, triplet loss embedding, blood pressure, prosody, triage session

1. INTRODUCTION

There has been a wide variety of research exploring the relationship between human physiological signals and their psychological states. In specifics, many have examined the interaction between the central cardiovascular and pain modulation system. For example, researchers have pointed out a phenomenon known as hypertension-induced hypoalgesia, which refers to a condition where patients with arterial hypertension perceive less pain and have lower pain sensitivity than normal individuals [1, 2]. Furthermore, postoperative chest pain experienced during physical exercise has also been shown to be inversely correlated with blood pressure (BP) [3]

Recent research further suggests that BP variation could be a more generalized indicator of the overall negative emotion experience and reactivity. For example, the subtle change in the central nervous system (CNS) has been shown to accompany or precede an increase in BP [4]. Functionally, blood pressure-related antinociception may represent a complex coordinated adaptive response of the body to stressful situations. Experiments have shown that higher BP is associated with dampened responses to negative emotional stimuli [5, 6], while other research found that BP can also damp responses to positively valenced stimuli [7]. Furthermore, research has also observed the connection between BP and several general affective traits, e.g., an influence of stress on BP variation [8], and the effect of inhibition of expressive negative emotion on BP changes [9]. Not only does the physiological signal provide a hint related to our inner state, it is a measure of our health condition regularly used in medical settings, e.g., BP is measured for every on-boarding emergency patient.

Few but recent research have started to indicate that changes in physiology are reflected in an individual's expressive vocal cues. For example, rapid rises in loudness and tempo are related to increases in BP [10]. Tsiartas et al. demonstrate that the changes in acoustic features are predictive of changes in speaker's HRV when interacting with a frustration-induced dialog agent [11]; at the same time, a BP evaluation method is proposed showing that it would be preferable to estimate BP by using voice-spectrum analysis [12]. Although detecting physiology changes from vocal cues have been studied, none of the works include real-world and clinically-spontaneous data collection in a medical setting. Further, there remains a lack of any sound learning framework in this domain. The ability to model the changes of physiological state from easily-obtainable vocal cues can open up opportunities for just-in-time patient monitoring.

In this work, we leverage a large-scale real patients database, i.e., collected at the Chang Gung Memorial Hospital for studies of pain in triage classification [13, 14], to develop a computational framework to predict changes in BP between *pre-* and *post-* triage sessions of on-boarding patients using prosodic cues. We propose a triplet-loss deep regressor network to infer changes of BP from prosodic features. The framework achieves a predicted correlation of 0.419 and 0.386 for SBP (systolic blood pressure) and DBP (diastolic

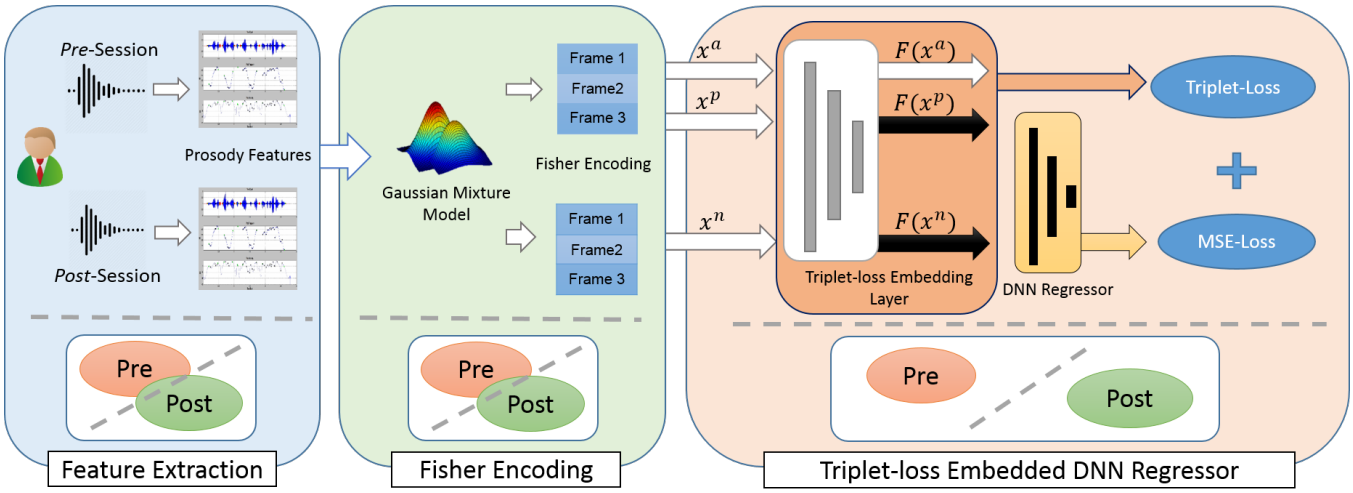

Fig. 1. Our proposed Fisher-vector based deep triplet-loss embedded deep regressor network to predict changes in blood pressure between *pre-* and *post-* intervention at emergency triage using prosodic features.

blood pressure), which is 26.1% and 17.3% relative improvement over DNN-regressor without a triplet-loss embedding layer. The use of a triplet-loss embedding layer helps mitigate issues of individual idiosyncrasy and is naturally appealing and applicable in individualized health monitoring.

The rest of paper is organized as follows: Section 2 describes our proposed framework. Section 3 shows our results and discussions. Section 4 is the conclusion.

2. RESEARCH METHODOLOGY

2.1. Chang Gung Audio-Video Pain Database

We use the database collected at the Chang Gung Memorial Hospital emergency department[1]. The database was originally collected to develop automated computational models to measure an on-boarding emergency room patient's pain intensity level during triage sessions. The database includes audio-video data (recorded using a Sony HDR handy cam in a fixed assessment room), measures of physiological data (body temperature, heart rate, respiration rate and blood pressure), numerical rating scale of pain intensity, and other clinical outcomes (analgesic prescription and patient disposition). Every patient is recorded at two points in time, i.e., an initial triage session (*pre-*) and a follow up session (*post-*) Each session lasts about 30 seconds to 1 minute. There are a total of 262 unique patients collected in the Chang Gung Audio-Video Pain database. Due to missing data and bad recording conditions, we use a subset of 94 unique patients, each collected at both *pre-* and *post-* session resulting in a total of 188 samples in this work. All patients' utterances are manually segemented, and staffs' voicing portions are excluded in this work. This constitutes the dataset used in this work and is one of the largest spontaneous real spoken datasets collected in a hospital setting.

[1]IRB#:CM104-3625B

2.2. Computational Framework

We will elaborate our proposed Fisher-vector based triplet-loss embedded deep regressor network in this section. Figure 1 depicts our overall framework, including prosodic low-level descriptors, Fisher-vector encoding, and triplet-loss embedded deep regressor network.

2.2.1. *Prosodic Low Level Features*

The Chang Gung Audio-Video Pain database has been pre-segmented into utterances with speaker identification marked. Previous studies on this database have indicated that prosodic features can be robustly extracted for the task of estimating pain [14]. Hence, in this work, we extract 9 low-level prosodic descriptors in total, including 1 pitch, 1 intensity, 1 harmonic-to-noise ratio and their associated delta and delta-delta every 10ms using the Praat toolkit [15]. These prosodic features are further z-normalized per speaker.

2.2.2. *Fisher Vector Encoding*

Fisher vector encoding approach projects the original low-level descriptors (LLD) feature space into a generative statistical representation with discriminative power [16]. The use of Fisher vector encoding originated from computer vision task and has since been recently proposed in the automatic paralinguistic analyses from voice [17]. We use Fisher vector encoding as our feature representation at the frame level. A brief description of Fisher vector encoding is below:

The low-level acoustic feature set is denoted by $X = \{x_t, t = 1...T\}$ with D dimensions, and the set of parameters of the GMM is $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1...K\}$ where w_i, μ_i and Σ_i corresponds to the zeroth (weight), first (mean vector) and second (covariance matrix) statistics for each mixture of Gaussian, respectively. The likelihood $p(x_t|\lambda)$ and posterior

Table 1. A summary of prediction results. Functional: 15 statistic values of low-level descriptors; FV: session-level Fisher-vector encoding; Regressor: a 4-layer DNN regressor; Triplet Regressor: proposed architecture with triplet layer embedding.

Targets	Pitch Functional	Intensity Functional	HNR Functional	Best Functional	Best Fisher Vector	Best Neural Network	
						Regressor	Triplet-Regressor
SBP	0.082	0.222	0.259	0.260 (PIH)	0.241 (PIH)	0.306 (PIH)	**0.386 (PIH)**
DBP	0.28	0.341	0.287	0.354 (PI)	0.369 (PI)	0.357 (PI)	**0.419 (PI)**

$\gamma_t(i)$ can be then computed using estimated μ_i and Σ_i,

$$p(x_t|\lambda) = \sum_{i=1}^{K} w_i p_i(x_t|\lambda) \quad (1)$$

$$\gamma_t(i) = p(i|x_t, \lambda) = \frac{w_i p_i(x_t|\lambda)}{\sum_{j=1}^{N} w_j p_j(x_t|\lambda)} \quad (2)$$

Then, the gradient functions can be computed below indicating the direction of movement in the GMM parameter space to properly fit the observed data sample.

$$g_{\mu_i}^X = \frac{1}{T\sqrt{\pi_c}} \sum_{t=1}^{T} r_t(k) \left(\frac{x_t - \mu_i}{\sigma_i} \right) \quad (3)$$

$$g_{\sigma_i}^X = \frac{1}{T\sqrt{2\pi_c}} \sum_{t=1}^{T} r_t(c) \left(\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right) \quad (4)$$

In this work, we transform the raw prosodic LLDs into Fisher vectors by concatenating $g_{\mu_i}^X$ and $g_{\sigma_i}^X$ computed every frame.

2.2.3. Triplet-loss Embedded Deep Regressor Network

The objective of this work is to estimate the change in BP of a patient between two assessment time (*pre*: triage session and *post*: follow-up session) from prosodic Fisher vector representation. A conventional approach in predicting the *change* is to the regress on the feature *difference*, e.g., Tsiartas et al. predict the difference in HRV using features by subtracting the acoustic descriptors from post-stimuli session to pre-stimuli session [11]. In this work, we propose a neural network architecture that combines a triplet-loss embedding layer to a standard deep neural network regressor. Triplet embedding has recently being applied for learning image descriptors and speaker-turn embedding [18, 19]. In a triplet-loss network, the inputs are a batch of triplet units $< x_i^a, x_i^p, x_i^n >$, where x_i^a and x_i^p belong to the same identity while x_i^a and x_i^n refer to the different identities. In our task, x_i^a and x_i^p is defined as patient's *pre*-session frame and x_i^n corresponds to the *post*-session. Let $f(x)$ denote the networks feature representation of input x which embeds the prosodic representation into a $d-$dimensional Euclidian space. For a training triplet $< x_i^a, x_i^p, x_i^n >$, the ideal feature representation should satisfy the following constraint:

$$\|f(x^a) - f(x^p)\| + \tau \leq \|f(x^a) - f(x^n)\| \quad (5)$$

where τ is a hyperparameter defining the minimum margin between the *pre* and the *post*-session. Thus, the triplet-loss function (L_{Triplet}) being minimized is defined below:

$$\sum_{1}^{N} \max\{\|f(x^a) - f(x^p)\| + \tau - \|f(x^a) - f(x^n)\|, 0\} \quad (6)$$

As shown in Figure 1, we construct layers of fully-connected network as triplet-loss embedding layer before feeding into the deep neural network regressors. The complete triplet-loss embedded deep regressor network is optimized using the following total loss function,

$$L_{\text{Total}} = L_{\text{MSE}} + \alpha * L_{\text{Triplet}} \quad (7)$$

where L_{MSE} is the mean square error to the target label and α refers to the weighting between the two losses.

2.2.4. Target Label Definition

The target label, $y(i)$, is defined for each patient i as:

$$y(i) = (\text{BP}_{pre}(i) - \text{BP}_{post}(i)) / \text{BP}_{pre}(i) \quad (8)$$

which is the percentage change of BP from triage (*pre*) to follow-up (*post*) session.

3. EXPERIMENTAL SETUP AND RESULT

3.1. Experimental Setup

The exact architecture of our triplet-loss embedded deep regressor network includes: the DNN regressor is composed of 4 fully-connected layers, and the dimensions are (2M)-M-10-1, where M represents the dimension of the Fisher vector input, the triplet embedding network includes 2 additional fully-connected layers with node number (2M)-M. The mixture number used in the GMM for Fisher vector encoding is selected empirically (grid search within the range $\{2, 4, 8, 16, 32, 64\}$). To train the triplet-loss network, we randomly sample frames in the two sessions for a patient forming the triplet pair $< x_i^a, x_i^p, x_i^n >$.

We carry out a 20-fold speaker-independent cross-validation for every experiment. The complete network is trained using Adam ($lr = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-08$), and the evaluation metric used is the Spearman correlation.

Table 2. The table lists top 10 correlated prosodic features to the target label. The superscript refers to 1st or 2nd order delta.

SBP Functional LLD	Corr.
Intensity^2-F12	**0.235216**
Intensity-F15	0.223238
Intensity-F11	0.218036
Intensity^1-12	0.202445
HNR^2-F8	0.201419
HNR-F11	0.200282
Intensity-F13	0.19785
HNR^1-F12	0.171682
HNR^2-F10	0.160296
Intensity^1-F14	0.155513
Top Features:	Intensity(*6), HNR(*4)

DBP Functional LLD	Corr.
Pitch-F3	**0.296825**
Intensity^1-F14	0.262962
Pitch-F10	0.262912
Intensity-F11	0.238102
Intensity-F2	0.228623
Intensity-F15	0.223016
HNR^1-F3	0.177839
Pitch-F14	0.166661
Intensity-F13	0.164762
Pitch-F1	0.160998
Top Features:	Pitch(*4), Intensity(*5)

3.2. Experimental Results and Analysis

Table 1 summarizes results of our prediction experiments on SBP and DBP using prosodic features. The columns titled 'Functional' refer to using 15 statistical properties (max, min, mean, median, std, 1 percentile, 99 percentile, 99 percentile - 1 percentile, skewness, kurtosis, min position, max position, 25 percentile, 75 percentile, 75 percentile - 25 percentile, denoted as F1-F15) calculated for each descriptor at the session-level, then the subtraction between *post*-session and *pre-session* feature vector is used to train the support vector regression to predict change in BP. The column of 'Fisher Vector' is done by performing session-level Fisher vector encoding, then the subtraction output between the two sessions is then fed through support vector regression

Our proposed network obtains the best accuracy of 0.386 (p-value $\sim$ 0.001) in predicting changes of SBP and 0.419 (p-value $<$ 0.001) in predicting changes of DBP. The use of triplet-loss embedded layer provides essential improvement when comparing to straightforward DNN regressor, in specific, it obtains a 26.1% and 17.4% relative gain for SBP and DBP respectively. Furthermore, the use of triplet-loss network consistently outperforms methods which uses feature subtraction between sessions, while the approach of feature subtraction can help in cases of estimating within-speaker physiology variation, the use of triplet-loss embedding evidently provides improved modeling power to handle the complexity of this non-linear individual idiosyncrasy.

3.2.1. Prosodic Feature Analysis

To further understand the relationship between each type of prosodic features and BP (SBP and DBP) changes, we compute the correlation of each functional feature to our target labels. The top-10 correlated features over 20 cross-validation folds are listed in Table 2. Generally, we observe that the difference in the characteristics of voice intensity and HNR is associated with the change in SBP, while the difference in the characteristics of pitch and voice intensity is associated with the change in DBP. This result also corroborates with the accuracies obtained in Table 1, i.e., the best accuracy achieved in predicting SBP change is by using pitch, intensity, and HNR (PIH), and the best accuracy obtained for predicting DBP change is by using pitch and intensity only (PI). Furthermore, we see that the temporal changes (deltas) of LLDs seem to be more indicative about the change in SBP than in DBP. While the exact mechanism of the change in BP in altering the manifested voice characteristics needs to be further studied, we present an initial investigation in this work.

4. CONCLUSION

In this work, we present a novel computational framework of triplet-loss embedded deep regressor network to predict the change in BP by modeling an individual's prosodic variations. We use a large-scale dataset collected in the real-world medical setting, in specifics triage session at the emergency department. Our proposed framework obtains significant predictive ability of the change in BP between two different time points, i.e., triage session and follow-up session, using the patient's prosodic characteristics. An analysis on the prosodic features reveals that all pitch, HNR, and intensity features can be indicative of the change in SBP of a patient, and pitch and intensity are associated with the change in DBP.

There are multiple future directions. One immediate direction is to incorporate facial expressions as additional easily-obtainable expressive behavior modalities with temporal modeling technique in the task of estimating inner physiological states changes from external behaviors to further improve our computational framework. Second, the deeper understanding of the relationship between the manifested prosodic variation and the physiological state changes is important to bring quantitative insights about the underlying physio-psychological mechanism. Lastly, the use of triplet loss embedding is intuitively appealing in patient monitoring, where the individual idiosyncrasy continues to be a challenging computational problem hindering the effective tracking of individual health progression. As for currently, there is still room for improving prediction accuracy, we will continue to develop an advanced framework for a variety of human behavior modeling tasks, especially in the healthcare domain [20, 21].

5. REFERENCES

[1] Stephen Bruehl and Ok Yung Chung, "Interactions between the cardiovascular and pain regulatory systems: an updated review of mechanisms and possible alterations in chronic pain," *Neuroscience & Biobehavioral Reviews*, vol. 28, no. 4, pp. 395–414, 2004.

[2] GA Reyes del Paso and CM Perales Montilla, "Haemodialysis course is associated to changes in pain threshold and in the relations between arterial pressure and pain," *Nefrología*, vol. 31, no. 6, 2011.

[3] Blaine Ditto, Bianca D'Antono, Gilles Dupuis, and Denis Burelle, "Chest pain is inversely associated with blood pressure during exercise among individuals being assessed for coronary heart disease," *Psychophysiology*, vol. 44, no. 2, pp. 183–188, 2007.

[4] Sergio Ghione, "Hypertension-associated hypalgesia," *Hypertension*, vol. 28, no. 3, pp. 494–504, 1996.

[5] James A McCubbin, Marcellus M Merritt, John J Sollers III, Michele K Evans, Alan B Zonderman, Richard D Lane, and Julian F Thayer, "Cardiovascular emotional dampening: The relationship between blood pressure and recognition of emotion," *Psychosomatic medicine*, vol. 73, no. 9, pp. 743, 2011.

[6] Ivan Nyklíček, Ad JJM Vingerhoets, and Guus L Van Heck, "Elevated blood pressure and self-reported symptom complaints, daily hassles, and defensiveness," *International journal of behavioral medicine*, vol. 6, no. 2, pp. 177–189, 1999.

[7] Daniel Z Wilkinson and Christopher R France, "Attenuation of positive and negative affect in men and women at increased risk for hypertension: A function of endogenous barostimulation?," *Psychophysiology*, vol. 46, no. 1, pp. 114–121, 2009.

[8] Maxwell V Rainforth, Robert H Schneider, Sanford I Nidich, Carolyn Gaylord-King, John W Salerno, and James W Anderson, "Stress reduction programs in patients with elevated blood pressure: a systematic review and meta-analysis," *Current hypertension reports*, vol. 9, no. 6, pp. 520–528, 2007.

[9] Emily A Butler, Boris Egloff, Frank H Wlhelm, Nancy C Smith, Elizabeth A Erickson, and James J Gross, "The social consequences of expressive suppression.," *Emotion*, vol. 3, no. 1, pp. 48, 2003.

[10] Erika Friedmann, Sue A Thomas, Denise Kulick-Ciuffo, James I Lynch, and Masazumi Suginohara, "The effects of normal and rapid speech on blood pressure.," *Psychosomatic Medicine*, vol. 44, no. 6, pp. 545–553, 1982.

[11] Andreas Tsiartas, Andreas Kathol, Elizabeth Shriberg, Massimiliano de Zambotti, and Adrian Willoughby, "Prediction of heart rate changes from speech features during interaction with a misbehaving dialog system," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] Motoki Sakai, "Feasibility study on blood pressure estimations from voice spectrum analysis," *International Journal of Computer Applications*, vol. 109, no. 7, 2015.

[13] Fu-Sheng Tsai, Ya-Ling Hsu, Wei-Chen Chen, Yi-Ming Weng, Chip-Jin Ng, and Chi-Chun 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.

[14] Fu-Sheng Tsai, Yi-Ming Weng, Chip-Jin Ng, and Chi-Chun Lee, "Embedding stacked bottleneck vocal features within an lstm architecture for automatic pain level classification during emergency triage," in *ACII*, 2017.

[15] P Boersma and D Weenink, "Praat speech processing software," *Institute of Phonetics Sciences of the University of Amsterdam. http://www. praat. org*, 2001.

[16] Florent Perronnin and Christopher Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[17] Heysem Kaya and Alexey A Karpov, "Fusing acoustic feature representations for computational paralinguistics tasks," *Interspeech 2016*, pp. 2046–2050, 2016.

[18] BG Kumar, Gustavo Carneiro, Ian Reid, et al., "Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5385–5394.

[19] Hervé Bredin, "Tristounet: triplet loss for speaker turn embedding," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5430–5434.

[20] Shrikanth Narayanan and Panayiotis 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.

[21] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan, "Signal processing and machine learning for mental health research and clinical applications [perspectives]," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 196–195, 2017.