Interspeech 2018
2-6 September 2018, Hyderabad

Automatic Assessment of Individual Culture Attribute of Power Distance using a Social Context-Enhanced Prosodic Network Representation

Fu-Sheng Tsai[1,3], *Hao-Chun Yang*[1,3], *Wei-Wen Chang*[2], *Chi-Chun Lee*[1,3]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]International Human Resource Development, National Taiwan Normal University, Taiwan
[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

fstsai@gapp.nthu.edu.tw, hgy@gapp.nthu.edu.tw, changw@ntnu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

Culture is a collective social norm of human societies that often influences a person's values, thoughts, and social behaviors during interactions at an individual level. In this work, we present a computational analysis toward automatic assessing an individual's culture attribute of power distance, i.e., a measure of his/her belief about status, authority and power in organizations, by modeling their expressive prosodic structures during social encounters with people of different power status. Specifically, we propose a center-loss embedded network architecture to jointly consider the effect of social interaction contexts on individuals' prosodic manifestations in order to learn an enhanced representation for power distance recognition. Our proposed prosodic network achieves an overall accuracy of 78.6% in binary classification task of recognizing high versus low power distance. Our experiment demonstrates an improved discriminability (17.6% absolute improvement) over prosodic neural network without social context enhancement. Further visualization reveals that the diversity in the prosodic manifestation for individuals with low power distance seems to be higher than those of high power distance.

Index Terms: behavioral signal processing, prosody, center-loss embedding, culture attribute, power distance

1. Introduction

Culture is a societal phenomenon resulting as a collective social norm of human societies. It is an important core value in shaping an individual's expressed behaviors, group interaction dynamics, decisions toward life, and even personalities [1]. Studying culture at its most basic unit, i.e., at an *individual* level, is critical in understanding such a phenomenon [2, 3]. *Power Distance* is a cultural construct describing the extent to which power inequalities is viewed as natural in a society [4, 5], and this construct can be used in describing individual's belief about status, authority and power in organizations [6]. Individuals with *high* power distance value tend to legitimize the difference in decision-making between people in high and low power positions; on the other hand, individuals with *low* power distance value prefer more equal status and interactions. The effect in the difference of power distances in shaping individual behaviors during human interactions has been identified in different social contexts, e.g., power distance construct moderates cross-level leadership-employee interaction [7] and people with low power distance show less response to lower-level of voice compared to those with high power distance [6].

Spoken dialogs is the most natural form of human's daily communication. Research has indicated that an individual's prosodic structures during spoken dialogs are not only related to their internal states (mental states, emotions, mood, etc) but also influenced by the role or the status of the dialog partners [8, 9, 10]. The expressive aspects of speech prosody are conceptualized to contain two distinct processes: the involuntarily-controlled expressions of affect and intentionally-controlled attitudinal functions of social factor [11]. Mixdorff et al. further examine the discriminatory power of macro-prosodic parameters in differentiating different attitudinal expressions [8, 12].

While research has been conducted in studying prosodic parameters and its communicative functions in different cultures, e.g., Shochi et al. have investigated the role of prosodic parameters in inter-cultural (English, French, Japanese) perception of affect [13], limited work has studied culture-prosody relationship at the individual level (not as an entire society). In this work, we present a computational study in automatic assessing individual's *power distance* measure by modeling the subjects' prosody as they engage in a situational question-answering social settings. In this experiment, we collecte audio data where the subject is given seven interaction scenarios where in each they are asked to bring up a particular question to three people with different relative levels of power status (same, slightly higher, and significantly higher), i.e., the three different social settings. Every subject is also being assessed on an established power distance scale [14]. We further propose a social context-enhanced prosodic network (SC-ePN) representation in this study aiming at recognizing individual culture value of power distance. The SC-ePN learns an enhanced representation on prosodic contours by jointly optimizing the network with a center-loss criterion (enhancing intra-class compactness) computed across the three different social settings in order to help uncover the discriminative portion of prosodic structure for power distance recognition.

Our proposed SC-ePN representation obtains an average (over seven types of scenarios) of 78.6% accuracy in classifying an individual between high and low power distance. The use of center-loss joint optimization that considers the effect of different social context settings improved the recognition rate by 21.3% relative improvement over deep prosodic network (DPN) representation without center-loss embedding. By visualizing the learned representation of our SC-ePN, the use of center-loss not only increases the power distance discriminatory power by centering the prosodic representation, we also observe a pattern that individuals with *high* power distance tend to show less variability in their expressive prosodic features as compared to those with *low* power distance. The rest of the paper is organized as follows: Section 2 describes our data collection, social context-enhanced prosodic representation, and classifier setup. Section 3 includes experimental setups, results and discussions. Section 4 concludes with future work.

10.21437/Interspeech.2018-1523

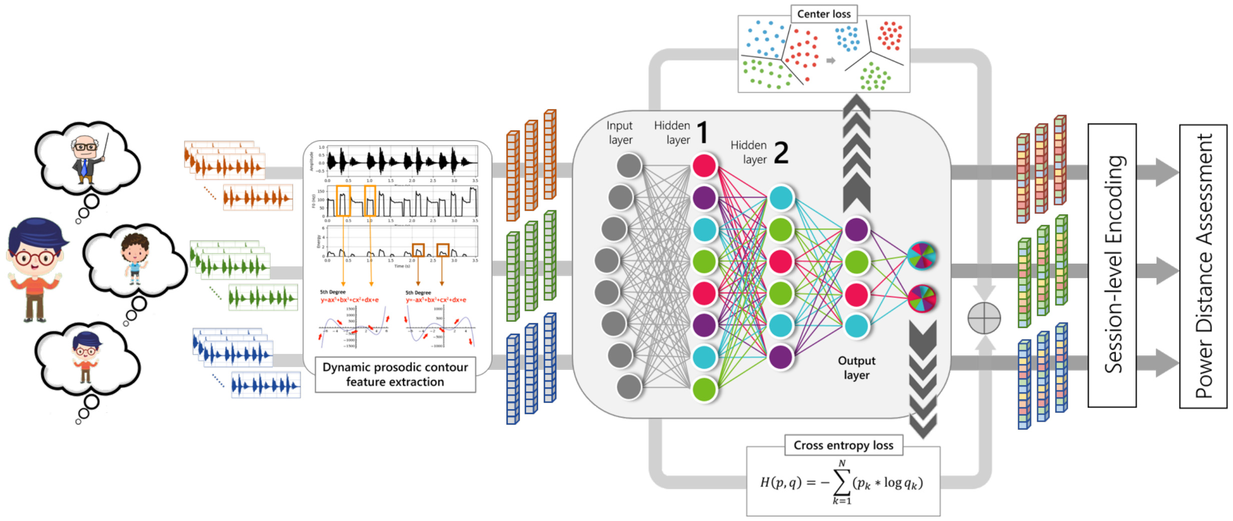

Figure 1: *It shows the complete architecture of our social context-enhanced prosodic network used for automatic power distance recognition: dynamic modeling of prosodic pitch and energy contour, training prosodic network by jointly optimizing setting-wise center-loss with standard cross entropy criteria, performing recognition using functional encoding of the network output layer with support vector classification.*

2. Research Methodology

2.1. Database Collection

In this work, we design an experimental protocol to study the relationship between prosodic manifestations and personal power distance cultural attribute. The experimental protocol asks the subject to first imagine a social setting where they have to ask a person of different social status to do something for them. Then one of our experimenter would play the role of this 'person' to engage in dialog with the subject. There are a total of seven questions and three social settings for each subject; every question-social setting pair constitutes an interaction scenario. The following is the list of seven conversation topics:

1. There are two free meal tickets, how would you invite him/her to join you?
2. You don't know how to do your homework, how would you ask him/her for help?
3. There is a job interview next week and you want some advice, how would you say to him/her?
4. After you had a fight with someone, you want to seek advice in handling the aftermath. How would you request for help?
5. Your family ran into financial difficulties and you are considering about quitting school to find a full-time job, how would you seek advice with him/her?
6. You failed the class with 3 points short preventing you from obtaining the final graduation credits, how would you ask for more points from him/her?
7. Your graduation exhibition would take place next week, how would you invite him/her to attend?

The three different social settings are: your respected professor or teacher, senior classmate/TA, and self-conversation. We recruit 26 participants in our experiment resulting in a total number of 546 ($26 * 3 * 7$) interaction scenarios in this dataset.

2.1.1. *Power Distance Scale (POW)*

The aim of our study is to automatically assess an individual's power distance using prosodic features. The main construct of power distance was developed by Hofstede [5] describing the extent to which power inequalities is viewed as natural in a society [4]. Sharma [14] demonstrated that power distance includes two dimensions: power (POW) and social inequality (IEQ). POW indicates how individuals are related to authority while IEQ shows one's hierarchical or egalitarian orientation. In this work, we use the POW scale derived from Sharma as the cultural measure of the subject's perception regarding authority and interactions in power relations. We further binarize the POW scale in our dataset, where 13 of them are considered as *high* POW, and 13 of them are assigned to *low* POW.

2.2. Social Context-Enhanced Prosodic Network

Figure 1 depicts a schematic of our complete proposed social context-enhanced prosodic network (SC-ePN). The SC-ePN is learned by introducing the use of center-loss embedding joint optimizing over social settings. We will describe the prosodic parameters and network architecture in the following.

2.2.1. *Acoustic Dynamic Prosodic Parameters*

Prosodic intonation structure is a key component in conveying social attitudes in Mandarin [15], and many research has also been conducted to understand aspects of attitudinal prosody in different languages [16, 17, 18]. In this work, we extract the following 13 dynamic prosodic features (30ms window size with 10ms step size) from the subject's speech during each interaction scenario.

- 1 Duration of the voice segment
- 6 Coefficients of 5-degree polynomial function to model pitch contour
- 6 Coefficients of 5-degree polynomial function to model energy contour

We further perform context window expansion to obtain a total of 39 features per frame. These prosodic features are z-normalized with respect to each speaker.

2.2.2. *Center-loss Embedded Network*

Prosodic structures parametrized by the dynamic acoustic parameters (section 2.2.1) include a vast amount of diverse infor-

Table 1: *It summarizes the Unweighted Averaged Recall (UAR) obtained in our proposed power distance recognition experiment. LLDs indicates the acoustic-prosodic low-level descriptors, DPN indicates representation derived from feed-forward neural network without center-loss, and SC-ePN is our proposed social context-enhanced prosodic network. C denotes three different kinds of social context settings, C1, C2 and C3 indicate professor, classmate and self.*

	LLDs	DPN	SC-ePN	LLDs	DPN	SC-ePN	LLDs	DPN	SC-ePN	LLDs	DPN	SC-ePN
	C1	C1	C1	C2	C2	C2	C3	C3	C3	Voting	Voting	Voting
Q1	73.1	65.4	69.2	69.2	73.1	69.2	57.7	42.3	69.2	**76.9**	65.4	73.1
Q2	57.7	53.8	**84.6**	46.2	61.5	76.9	57.7	57.7	80.8	50.0	69.2	**84.6**
Q3	73.1	61.5	65.4	65.4	53.8	**80.8**	69.2	50.0	76.9	76.9	65.4	73.1
Q4	53.8	50.0	**76.9**	**76.9**	46.2	73.1	53.8	53.8	69.2	65.4	50.0	**76.9**
Q5	42.3	57.7	80.8	65.4	57.7	65.4	57.7	65.4	**84.6**	57.7	57.7	**84.6**
Q6	50.0	80.8	69.2	88.5	84.6	69.2	76.9	57.7	65.4	**92.3**	84.6	69.2
Q7	53.8	57.7	76.9	57.7	76.9	**92.3**	42.3	38.5	76.9	46.2	61.5	88.5
Avg.	57.7	61.0	74.7	67.0	64.8	75.3	59.3	52.2	74.7	66.5	64.8	***78.6***

mation, e.g., emotion, attitude, situation, etc. Directly learning a classifier from these parameters is inadequate to uncover the fine-grained culture value of power distance. We propose to enhance the discriminatory power in these representations using a a novel prosodic network architecture (SC-ePN) that simultaneously considers the prosodic structures in the three different social settings using center-loss embedding. The use of center-loss embedding has recently been applied for recognition tasks, for instance, face recognition [19], person recognition [20], and handwritten Chinese character recognition [21], etc. The center-loss function being minimized is defined below:

$$L_c = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2 \tag{1}$$

where m is the number of training samples in a batch. x_i is the i^{th} training sample. y_i is the class (social setting) corresponding to x_i. c_{y_i} donates the y_i^{th} class center.

The SC-ePN includes three layers of fully-connected layers (Figure 1). The complete loss function in learning SC-ePN is a combination of center-loss, L_c, that learns a centralized setting-specific feature space and the target label loss, L_{CE}, that learns to classify between high POW versus low POW.

$$L_{Total} = L_{CE} + \lambda L_c \tag{2}$$

where L_{CE} is the cross entropy to the target label and λ refers to the weighting between the two losses (we set 0.5 in this work).

2.3. Power Distance Classification

The SC-ePN outputs frame-level prosodic representation for every subject's interaction scenario. Since every session is of different length, it results in varying number of sequences. We additionally compute 15 statistical functionals to generate the final feature vector of each participant's session-level feature vector inputted to the classifier. The list of functionals included maximum, minimum, median, mean, standard deviation, 1st percentile, 99th percentile, 99th -1st percentile, skewness, kurtosis, maximum position, minimum position, upper quartile, lower quartile and interquartile range. The selected classifier for training and recognition is linear-kernel support vector machine.

3. Experimental Setup and Results

We report recognition results on binary classification between high and low power distance. Accuracy was measured in unweighted average recall (UAR) with the evaluation scheme done via leave-one-person-out cross-validation.

3.1. Experimental Setup

The SC-ePN architecture is composed of 3 fully-connected layers with the node dimensions at every layer to be 39-16-2. The total loss function is composed of center loss for social settings and cross entropy loss for the target power distance label. The batch size, epoch and iteration in epoch are set at 500, 5 and 1000, respectively. The complete network is trained using Adam ($lr = 0.001$). The output layer (16 dimensions) is extracted as participant's acoustic representation at the frame-level. We compare our method to the following two other methods:

- *LLDs:* Compute 39 dynamic prosodic features as the participant's frame-level prosodic representation.
- *DPN:* Learn a network with the same structure as SC-ePN without center-loss embedding in deriving the frame-level prosodic representation.

These features are then fed into statistical function-based session-level encoding to perform final social power distance recognition for each participant.

3.2. Experimental Results and Discussions

Table 1 summarizes our complete experimental results. C denotes the three different kinds of of social settings for each subject (C1: professor, C2: classmate/TA, C3: self); LLDs and DPN are our baseline methods mentioned in section 3.1; Q1-7 are the seven questions/topics that each participant engage in as mentioned in section 2.1; SC-ePN denotes our proposed social context-enhanced prosodic network architecture; Avg. denotes the averaged recognition UAR of all seven scenarios.

The SC-ePN achieves the best recognition rates among the three social settings, especially in the C2 setting; it obtains a 75.3% recognition rate compared to LLDs (67.0%) and DPN (64.8%), i.e., 12.4% and 21.3% relative improvement, respectively. Since an individual only has a single power distance measure, we can perform majority vote over the three settings, denoted as "XX-Voting" in Table 1. The SC-ePN Voting obtains a further improvement in the average recognition rates of 78.6% (18.2% and 13.8% relative improvement over LLDs and DPN voting strategy, respectively). In overall view, the Avg. is much more interesting that the relative improvement between SC-ePN and the other two methods (LLDs and DPN) is much larger for C1 and C3. Our experiment clearly indicates that through joint optimization using center-loss embedded network over social settings provided the desirable approach for social-context feature embedding and certainly essential improvement. This may be attributed to the computational effect of non-linear centralization of prosodic feature space within each social setting, it effectively uncovers the discriminative portion within prosodic structure for cultural value of power distance.

In general, we observe that the proposed SC-ePN-Voting structure provides promising modeling power to assess an individual's culture trait of power distance. However, we observes that for Q1 and Q6 (especially Q6), the use of center-loss embedding negatively impacts the recognition results; the original LLDs method achieves fairly high accuracy along. We would like to further investigate whether the design of question/topic would have an effect in eliciting the prosodic variations used in this context of culture trait recognition.

3.2.1. SC-ePN Visualization

An analysis on visualizing the difference between the learned SC-ePN features and the DPN features is presented. Figure 2 (left) shows an example of two subjects prosodic representations using DPN in Q5, and Figure 2 (right) shows the same two subjects prosodic representing using SC-ePN (both of them are visualized using t-SNE [22]) . The blue dots indicates data from the subject of *high* power distance, and the red dots indicates data from the subject of *low* power distance. Here, we define the average intra-class distance $D(PDI)$ as below:

$$D(PDI) = \frac{\sum_{i=1}^{N_p} \|x_i - c\|_2^2}{N_p} \quad (3)$$

where N_p is the number of data from class p , x_i is data sample and c is the center of data from class p. This measure quantifies the spread of the prosodic variation after the network attempt to centralize their representation (noted in Figure 2).

By observing Figure 2, we can see the effect in using center-loss in this context. Originally, the feature space, while discriminatively trained using cross-entropy loss computed with respect to the target label, it is still highly-overlapping between subjects of high vs. low power distance. However, after introducing center-loss with respect to the social setting, the prosodic features of either high or low power distance subject become highly concentrated and clearly non-overlapping. Another interesting observation that we observe from visualization figures is that its prosodic variation for *low* power distance is higher than for *high* distance subject, and the average $D(PDI)$ for *high* and *low* power distance in our database are 17.685 and 18.119, respectively.

4. Conclusions

In this work, we present an initial study into automatic assessment of an individual's culture trait of power distance by modeling their expressive prosodic structures. We design an experiment where the subject engage in a set of scenarios with partner of different social status. We further propose a social context-enhanced prosodic network (SC-ePN) to obtain a promising recognition accuracy in assessing trait of power distance. Our SC-ePN provides an enhanced prosodic representation by jointly considering social settings with a center-loss criterion in the network training. Our analyses reveal that by centralizing prosodic representation with respect to each social setting, our SC-ePN effectively helps uncover the discriminative portion of prosodic structure for power distance recognition. We further observe a phenomenon that under these social interaction scenarios, the diversity in the prosodic variations for individuals with high power distance seems to be less than those of low power distance.

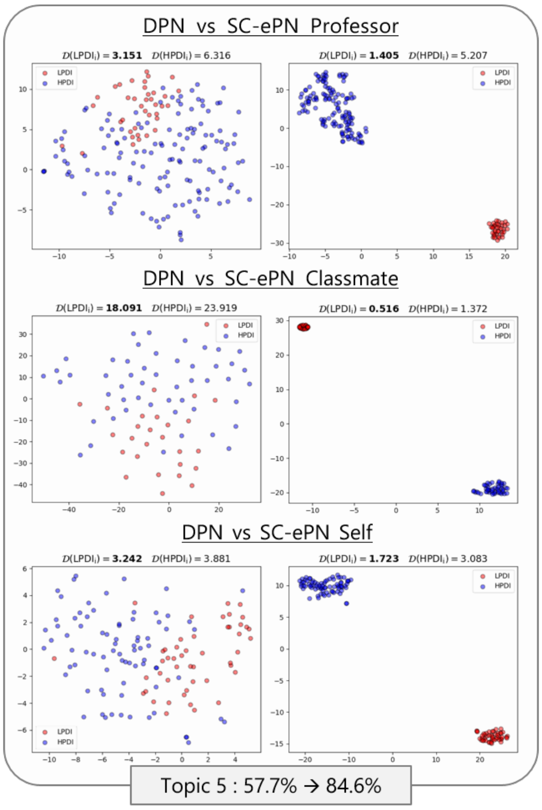

Figure 2: *A visualization analysis of the prosodic network representation learned with center-loss (right side) and without center-loss (left side). The D(PDI) indicates the averaged distance from all data samples to class center. Blue dots indicates data samples from high power distance subject, and Red dots indicates data samples from low power distance subject.*

There are several future works to pursue. One immediate work is to expand the scale and the subject numbers of the current database to provide more robust insights and conclusions on relating expressive prosodic structures and personal culture trait of power distance. Secondly, aside from prosodic cues, we would like to investigate and jointly model the subject's non-verbal behaviors (e.g., body gestures, facial expressions, even head orientation) during social interaction scenarios to achieve higher recognition rates. Furthermore, as part of the study, we have already collected brain images for the same set of subjects in a similar experimental protocol, i.e., they are exposed to different social scenarios with people of different social status inside the magnetic resonance imaging scanner. By cross-referencing internal brain responses, expressive verbal/non-verbal behaviors, and validated self-assessments of an individual cultural values, we will continue to advance our technical framework in quantitative understanding the influence of personal culture value on their expressive behaviors to inspire further development of behavior analytic for tasks of human-centered applications and research [23].

5. References

[1] R. Linton, *The cultural background of personality*. New York ; London : D. Appleton-Century company, incorporated, 1945, five lectures delivered at Swarthmore college under the auspices of the Cooper foundation, February, 1943. cf. Pref.

[2] E. Karahanna, J. R. Evaristo, and M. Srite, "Levels of culture and individual behavior: An integrative perspective," *Advanced Topics in Global Information Management*, vol. 5, no. 1, pp. 30–50, 2006.

[3] H. C. Triandis, "The self and social behavior in differing cultural contexts." *Psychological review*, vol. 96, p. 506, 1989.

[4] J. Brockner, G. Ackerman, J. Greenberg, M. J. Gelfand, A. M. Francesco, Z. X. Chen, K. Leung, G. Bierbrauer, C. Gomez, B. L. Kirkman *et al.*, "Culture and procedural justice: The influence of power distance on reactions to voice," *Journal of Experimental Social Psychology*, vol. 37, no. 4, pp. 300–315, 2001.

[5] G. Hofstede, *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2003.

[6] B. L. Kirkman, G. Chen, J.-L. Farh, Z. X. Chen, and K. B. Lowe, "Individual power distance orientation and follower reactions to transformational leaders: A cross-level, cross-cultural examination," *Academy of Management Journal*, vol. 52, no. 4, pp. 744–764, 2009.

[7] R. Loi, L. W. Lam, and K. W. Chan, "Coping with job insecurity: The role of procedural justice, ethical leadership and power distance orientation," *Journal of Business Ethics*, vol. 108, no. 3, pp. 361–372, 2012.

[8] H. Mixdorff, A. Hönemann, and A. Rilliard, "Acoustic-prosodic analysis of attitudinal expressions in german," *Proceedings of Interspeech 2015*, 2015.

[9] R. Van Bezooijen, "Quality for the attribution of social status and personality characteristics," *Language attitudes in the Dutch language area*, vol. 5, p. 85, 1988.

[10] D. S. Hurley, "Issues in teaching pragmatics, prosody, and non-verbal communication," *Applied Linguistics*, vol. 13, no. 3, pp. 259–280, 1992.

[11] V. Aubergé, "A gestalt morphology of prosody directed by functions: the example of a step by step model developed at icp," in *Speech Prosody 2002, International Conference*, 2002.

[12] A. Barbulescu, R. Ronfard, and G. Bailly, "Which prosodic features contribute to the recognition of dramatic attitudes?" *Speech Communication*, vol. 95, pp. 78–86, 2017.

[13] T. Shochi, A. Rilliard, and V. Aubergé, "Donna erickson intercultural perception of english, french and japanese social affective prosody," *The role of prosody in Affective Speech*, vol. 97, p. 31, 2009.

[14] P. Sharma, "Measuring personal cultural orientations: Scale development and validation," *Journal of the Academy of Marketing Science*, vol. 38, no. 6, pp. 787–806, 2010.

[15] W. Gu, T. Zhang, and H. Fujisaki, "Prosodic analysis and perception of mandarin utterances conveying attitudes," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[16] H. Fujisaki and K. Hirose, "Analysis and perception of intonation expressing paralinguistic information in spoken japanese," in *ESCA Workshop on Prosody*, 1993.

[17] D.-K. Mac, V. Aubergé, A. Rilliard, and E. Castelli, "Vietnamese multimodal social affects: How prosodic attitudes can be recognized and confused," in *Spoken Languages Technologies for Under-Resourced Languages*, 2010.

[18] T. Shochi, G. Gagnié, A. Rilliard, D. Erickson, and V. Aubergé, "Learning effect of prosodic social affects for japanese learners of french language," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[19] Y. Xu, H. Ma, L. Cao, H. Cao, Y. Zhai, V. Piuri, and F. Scotti, "Robust face recognition based on convolutional neural network," *DEStech Transactions on Computer Science and Engineering*, no. icmsie, 2017.

[20] Y. Liu, H. Li, and X. Wang, "Learning deep features via congenerous cosine loss for person recognition," *arXiv preprint arXiv:1702.06890*, 2017.

[21] R. Zhang, Q. Wang, and Y. Lu, "Combination of resnet and center loss based metric learning for handwritten chinese character recognition," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 5. IEEE, 2017, pp. 25–29.

[22] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[23] 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.