I-vector Based Physical Task Stress Detection with Different Fusion Strategies

Chunlei Zhang, Gang Liu, Chengzhu Yu, John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas, USA.
{chunlei.zhang, john.hansen}@utdallas.edu

Abstract

It is common for subjects to produce speech while performing a physical task where speech technology may be used. Variabilities are introduced to speech since physical task can influence human speech production. These variabilities degrade the performance of most speech systems. It is vital to detect speech under physical stress variabilities for subsequent algorithm processsing. This study presents a method for detecting physical task stress from speech. Inspired by the fact that i-vectors can generally model total factors from speech, a state-of-the-art i-vector framework is investigated with MFCCs and our previously formulated TEO-CB-Auto-Env features for neutral/physical task stress detection. Since MFCCs are derived from a linear speech production model and TEO-CB-Auto-Env features employ a nonlinear operator, these two features are believed to have complementary effects on physical task stress detection. Two alternative fusion strategies (feature-level and score-level fusion) are investigated to validate this hypothesis. Experiments over the UT-Scope Physical Corpus demonstrate that a relative accuracy gain of 2.68% is obtained when fusing different feature based i-vectors. An additional relative performance boost with of 6.52% in accuracy is achieved using score level fusion.

Index Terms: physical stress detection, i-vector, TEO-CB-Auto-Env, AdaBoost

1. Introduction

Stress is an external aspect that impacts physical speech production when people produce speech while performing secondary tasks. Addressing noise is not sufficient to overcome performance loss in actual noisy stressful scenarios for robust speech systems, even if noise is eliminated completely [1]. Speech production variability introduced by stress or emotion can severely degrade speech/speaker recognition accuracy [2-4]. Detection of paralinguistic information, such as physical task load, gender and cognitive load can guide human computer interaction systems to automatically understand and adapt to different users states and environments. Thus, this technique can be directly applied to stress level classification [5], as well as emotion surveillance. At the same time, it can also be employed as a front-end for spoken dialog systems, speaker diarization, speaker identification and automatic speech recognition (ASR) systems.

Table 1 shows an overview of factors which impact physical task stress detection. Physical status (e.g. heart rate) is changing with exertion, which can be reflected in the corresponding speech [6]. The acoustic environment or noise level varies with different physical task scenarios, which could be sustained background noise in a typical 24-hour operating workplace or random noise in a gym. Even if we remove all external environmental factors, physical task stress still shows differences within speaker (e.g. the same speaker, different exercise durations give different stress load) and across speaker (e.g. the same task, different speakers show different stress load levels). These factors taken together make physical stress detection a challenging research task.

Table 1: *Influential factors for speech under physical task.*

Physical Changes of Speaker	Noise type / Environment	Speaker Variability
● Heart rate	● Workplace	●Within speaker
● Breathing	● Gym	
● Fatigue	● Constant	●Across speaker
● Muscle control	● Random	

Due to importance of stress detection in real world speech applications, more attention has been drawn to this domain in the past decade [7,8]. Inspired by nonlinear speech production model, the Teager Energy Operator (TEO) based features have been well known to represent traits of stressed state by reflecting variability in the excitation [2,7]. Linear speech production model based features such as MFCCs are still effective since they reflect excitation characteristics. In [9], low-level descriptors (LLD) features employing the OpenSMILE extractor were generated as input features for a baseline system on a similar task data.

Speaker recognition systems based on i-vector extraction and PLDA classification are able to obtain relatively high accuracy and have become a mainstream framework for speaker identification tasks [10-14]. Inspired by the total variability modeling of speech, we focus on physical task stress detection using an i-vector framework. Performance on MFCCs and TEO-CB-Auto-Env features are explored. A fusion of these features at an i-vector level is considered to supplement performance, since MFCCs and TEO-CB-Auto-Env are derived from different speech production models. Finally, score fusion employing AdaBoost is also employed to provide further performance gain.

The remaining sections are organized as follows: Sec.2 presents a brief introduction to the physical task stress corpus used in the study. Sec.3 explains the i-vector framework implemented. Fusion at the i-vector and score level are presented in Sec.4. We report results and provide discussion in Sec.5. Conclusions and future work are explored in Sec.6.

2. UT-Scope Corpus

This study employs the UT-Scope Physical task stress Corpus

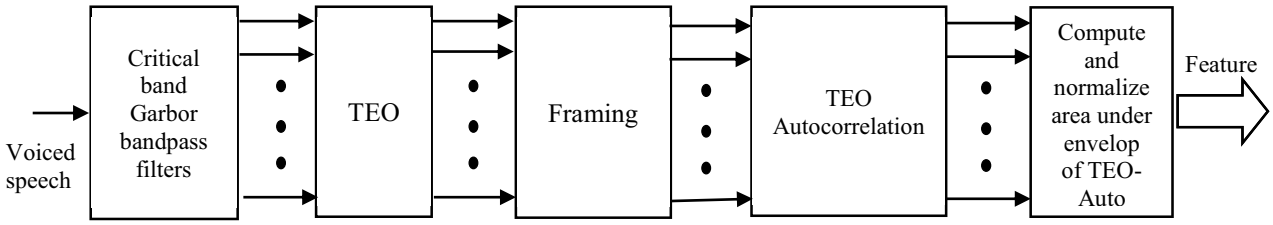

Figure 1: TEO-CB-Auto-Env feature extraction.

for system development and evaluation. From the corpus details described in [15], physical task stress is introduced into speech by having subjects exercise on a Stamina Conversion II Elliptical/Stepper machine in the elliptical mode. There are 66 speakers in the UT-Scope Physical Corpus. For this portion of the study, we employ 50 female speakers, each producing 35 sentences under neutral and physical conditions. We consider the stress classification experiments in a speaker independent scenario. In each experiment, 40 speakers are used in the training set, and 10 subjects in the test set. Next, we rotate the training/test set, resulting in 5 speaker independent physical stress detection experiments where all speech and speakers are open test. For more details, please see Table 2.

Table 2: *Statistics of UT-Scope Corpus.*

	Set1	Set2	Set3	Set4	Set5
SNR/dB	33.28	36.87	36.07	35.84	37.78
Duration/s	2.56	2.86	3.06	2.97	3.26
Spk Count	10	10	10	10	10

3. I-vector framework for stress detection

Our proposed system for physical task stress detection utilizes the concept of i-vector modeling, which is proposed in [10,16]. By constraining the total variability into a lower dimensional total variability space, the i-vector is capable of effectively representing the variability factors within each speech utterance [16,17]. In this work, we attempt to model the speaker-independent physical task stress using i-vectors. To compensate for MFCCs based i-vector, which is derived from linear speech production model, a nonlinear speech production model based frame-level feature entitled TEO-CB-Auto-Env [7] is investigated to extract utterance-level features.

3.1. Acoustic feature extraction

- MFCCs:

  39-dimention feature vector (13 MFCC+Δ+ΔΔ).

- TEO-CB-Auto-Env:

  The TEO profile obtained from the critical band based Gabor bandpass filter output is segmented on a short-term basis, Auto-correlation is applied after framing. Once the auto-correlation response is found, the area under the autocorrelation envelope is obtained and normalized. One area coefficient is obtained for each filter bank. This area coefficient is intended to determine the regularity of speech production, it has been shown to be large for neutral speech and low for speech produced under stressed conditions [7,18]. In this study, we employ an 18 dimensional Gabor filterbank. Thus, 18 dimensional TEO-CB-Auto-Env features are extracted from each frame. Fig.1 shows a flow diagram of the TEO-based feature extraction.

For each speaker independent experiment, a Universal Back-ground Model (UBM) with 256 Gaussian mixtures is trained using the training dataset outlined in Sec.2.

3.2. I-vector extraction

For utterance-level physical task stress detection, the i-vector modeling is given as:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}, \quad (1)$$

where **M** is the GMM supervector for an utterance, **m** is the stress-, utterance- and speaker-independent supervector obtained from UBM, **T** is the low rank total variability matrix representing the basis of the reduced total variability space, and **w** is the low rank factor loadings referred to as i-vector, which contain all factors related to physical stress and neutral conditions. The rank of **T** determines the dimension of the derived i-vector. In this work, 50, 100 and 200 dimensional i-vectors are examined because of the relatively short duration of each utterance (2-3s) [19]. Fig. 2 gives the flow chart of this i-vector framework.

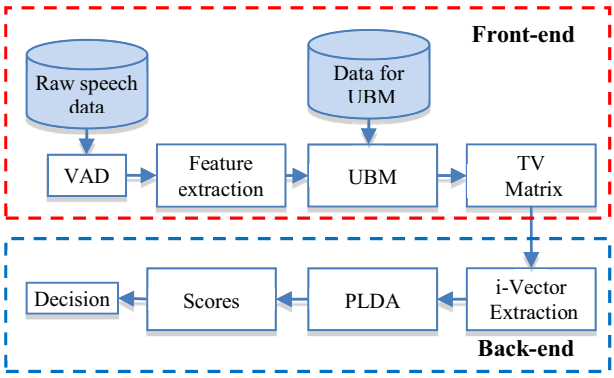

Figure 2: *Flow diagram of i-vector framework.*

4. Fusion in i-vector and score level

To fully leverage the physical/neutral discriminative information obtained from MFCCs and TEO-CB-Auto-Env features, system fusion is considered. Two different fusion strategies are examined in this section.

4.1. I-vector level fusion

Using the i-vector extraction described in Sec. 3, two kinds of i-vector are derived from each utterance, (i.e., MFCC-based and TEO-CB-Auto-Env based i-vectors). The new i-vector integrating both MFCC and TEO-CB-Auto-Env acoustic information

is obtained by concatenating the two i-vectors together. The dimensionality is reduced to the original length using linear discriminative analysis (LDA) [11].

4.2. Score level fusion

In practical binary classification as seen in Fig. 4, the decision is made by comparing the score difference of neutral (labeled with 1) and physical stressed speech (labeled with -1) with 0. Let us define S as the score difference, S_{neu} as the score of the neutral speech and S_{phy} as the score of physical stressed speech. Next, the decision is given by:

$$S = S_{neu} - S_{phy} : \begin{cases} \geq 0 & Neutral \\ < 0 & Physical \end{cases} \quad (2)$$

With the score difference *S*, a score map for the two features is illustrated as Fig. 3. From the score distribution, a new boundary could be learned which is beneficial for final decision making. Therefore, a score level fusion can be treated as a pattern classification problem. The scores obtained from different i-vector systems (i.e., TEO-CB-Auto-Env i-vector, MFCC i-vector, and i-vector fusion of two features) can be used as new feature vectors of both supervised classification algorithms (support vector machine (SVM), boosting, etc.) and unsupervised classification algorithms (clustering algorithms such as K-means) [20,21,22].

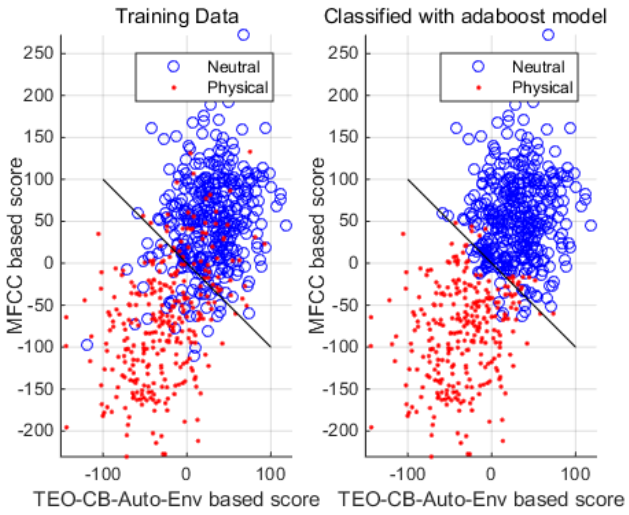

Figure 3: *Score map of neutral and physical stressed utterances. The line is the decision boundary.*

In our study, the AdaBoost algorithm is employed to learn the pattern of the score distribution [23]. The idea behind AdaBoost is to construct a strong classifier out of a set of weak classifiers. A final decision is then given by:

$$S(x_i) = \sum_{i=1}^{K} \alpha_i h_i(x_i), \quad (3)$$

where $h_i(x_i)$ is the weaker classifier of a given pattern x_i, the value for $h_i(x_i)$ is "yes" (+1) or "no" (-1) for binary classification, and α_i is the weight assigned to each classifier. When applying AdaBoost to our physical task stress detection system, we assume the MFCCs score S_1 and TEO-CB-Auto-Env score S_2 are presented by a weaker classifier respectively (Actually, we can claim additional weaker classifiers to present a two-dimension feature, since third or high number classifiers are just the linear combination of first two classifiers). By setting the threshold th_i to minimize the input training error, α_i is calculated according to the adapt rule of AdaBoost algorithm. Greater details concerning of AdaBoost can be found in [20,23]. In our proposed system, *K* in Fig. 4 is set to 2 as shown.

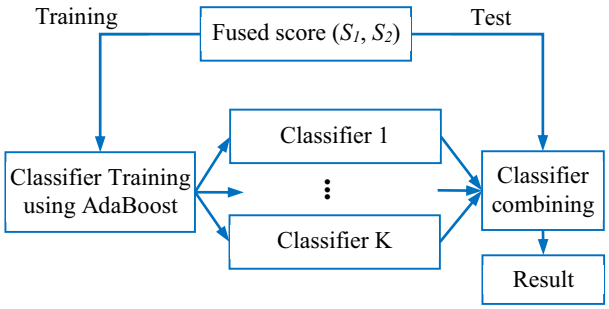

Figure 4: *Flow diagram of score fusion using AdaBoost.*

5. Experimental results

In this section, we present experimental results of physical task stress detection with our proposed systems. First, we examine the influence of different i-vector dimensions. Five-fold cross validation is employed as described in Sec. 2. The results shown in Fig. 5 represent the average over 5 test sets. 100 dimensional i-vector for all three features always outperform 50 dimensional i-vector since 100 dimensional i-vectors can carry more information than 50, however the difference is small. As i-vector dimension increases to 200, the performance does not always increase. In the following experiments, we set i-vector dimensionality to 100 for its relative stable performance.

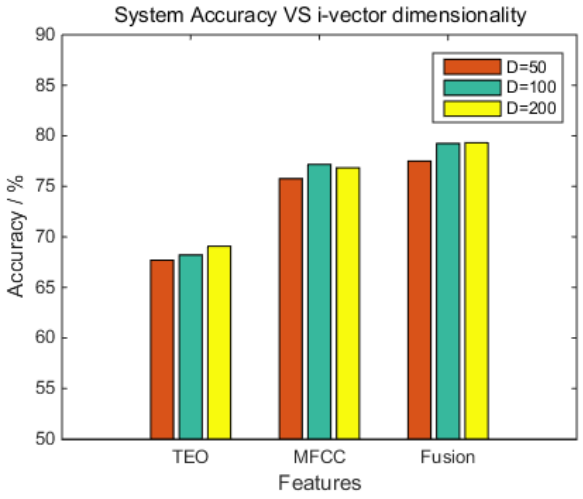

Figure 5: *Accuracy across different i-vector dimensions.*

In the score level fusion stage, a new score is given by (3), where $h_i(x_i)$ is defined as:

$$h_i(s_i) = \begin{cases} 1, & S_i > th_i \\ -1, & S_i < th_i \end{cases}, \quad (4)$$

In practice, we find that a weaker classifier such as Eq. (4) does not use the full information of score S_1 and S_2, where a hard decision boundary is obtained. From this observation, a "soft" boundary is given for practical classifiers.

$$S = sign\left[\sum_{i=1}^{2} \alpha_i S_i\right] \quad (5)$$

By this modification of the decision making rule, an approximately +1% accuracy gain is achieved. We set α_1 as 1.00 and α_2 as 0.55 (normalized by α_1) which is obtained from the AdaBoost training of scores.

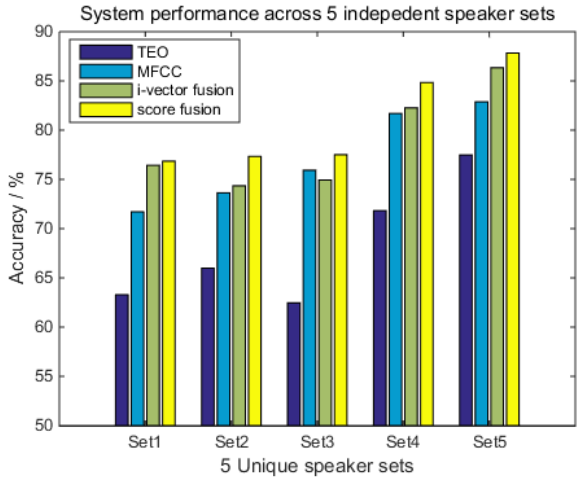

Figure 6: *System performance across 5 unique speaker set. We use scores from MFCC system and i-vector fusion system to perform score-level fusion.*

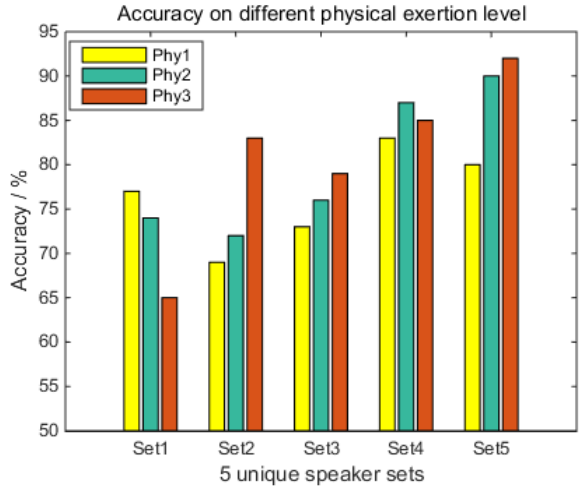

Figure 7: *Detection performance on physical stressed speech employing MFCC based i-vector system. As described in Sec. 2, each speaker produces 35 utterances under physical task condition. We split them into three categories (i.e. Phy1 for 1-12, Phy2 for 13-24 and Phy3 for 25-35).*

From the speaker independent physical stress detection experiments given in Fig. 6, we can see: a) both i-vector fusion and score fusion achieve reasonable performance, which show the effectiveness of MFCCs and TEO-CB-Auto-Env features and their complementary effects; b) compared to i-vector fusion, score fusion always performs better than single feature based systems, which shows the stability of our proposed approach; c) there is a greater than 10% percent accuracy difference between Set1 and Set5 indicates the variability across speakers; and d) although Set4 has lower SNR and shorter duration compared to Set2 and Set3 (see Table 2), the relative better detection performance further shows the across speaker variability of physical stress, which reflects a challenge in formulating a robust physical stress model.

To examine the physical exertion level reflected by speech within each speaker, we split each speaker set into 3 parts over the exercise fine frame (e.g., begin, middle, end) and repeat the physical stress detection experiments. We assume the physical stress load follows in this order: Phy3>Phy2>Phy1, since the entire exercise time period follows in the same way. The results from Fig. 7 show: a) physical stress load increases with the exercise time period, which indicates the variability introduced by physical exertion level within each speaker with exception of Set1, others generally show increasing physical stress level over time; b) the results here show that effective physical stress detection is possible, and increasing levels of stress are seen across the exercise period; c) it should also be noted that corresponding heartrate monitoring during speech production for the UT-Scope Physical Corpus collection confirm the increased levels of physical task stress [6].

6. Conclusions

In this study, an i-vector based physical task stress detection system was proposed. MFCCs and TEO-CB-Auto-Env based features were investigated in an i-vector framework for stress detection tasks. Using i-vector fusion, a relative accuracy gain of +2.68% is obtained; by score fusion using the AdaBoost algorithm, a further relative +6.52% performance gain is achieved (both compared to best single feature system used in our study, e.g. MFCC based i-vector system). The i-vector dimensionality for our specific physical task stress detection is determined by parameter tuning. Variability across and within speakers was investigated. From the experiments presented in Fig. 7, it has been shown that approximate physical exertion level differences are represented in the speech signal. Future work will focus on physical stress level classification, especially over speakers given heart rate ground truth. Also, other variations such as gender, channel or age will be explored.

7. Acknowledgments

This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

8. References

[1] J.H.L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication,* vol. 20, no. 1, pp: 151-173, 1996.

[2] J.H.L. Hansen, S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker Classification I.* Springer, 2007. pp. 108–137.

[3] J.H.L. Hansen, A. Sangwan, W. Kim, "Speech under stress and lombard effect: impact and solutions for forensic speaker

recognition," in *Forensic Speaker Recognition*. Springer, 2012, pp.103–123.
[4] C. Baber, B. Mellor, R. Graham, J.M. Noyes, C. Tunley, "Workload and the use of automatic speech recognition: The effects of time and resource demands," *Speech Communication*, vol. 20, no. 1, pp. 37–53, 1996.
[5] J.H.L. Hansen, E. Ruzanski, H. Bořil, J. Meyerhoff. "TEO-based speaker stress assessment using hybrid classification and tracking schemes." *International Journal of Speech Technology,* vol. 15, no. 3, pp: 295-311, 2012.
[6] K. W. Godin, J.H.L Hansen. "Analysis and perception of speech under physical task stress." In *Proc. INTERSPEECH*, 2008.
[7] G. Zhou, J.H.L. Hansen, J. F. Kaiser, "Nonlinear feature based classification of speech under stress*," IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.
[8] S.A. Patil, J.H.L. Hansen, "Detection of speech under physical stress: Model development, sensor selection, and feature fusion," in *Proc. INTERSPEECH*, 2008.
[9] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, Y. Zhang, "The Interspeech 2014 Computational Paralinguistics Challenge: Cognitive &Physical Load," in *Proc. INTERSPEECH*, 2014.
[10] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
[11] G. Liu, J.H.L. Hansen, "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1978–1992, 2014.
[12] C. Yu, G. Liu, S. Hahm, J.H.L. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," *IEEE. ICASSP*, 2014, pp. 4017–4021.
[13] C. Yu, G. Liu, J.H.L Hansen. "Acoustic Feature Transformation using UBM-based LDA for Speaker Recognition." in *Proc. INTERSPEECH*, Singapore, 2014.
[14] G. Liu, et al. "Investigating state-of-the-art speaker verification in the case of unlabeled development data." *Proc. Odyssey speaker and language recognition workshop*, Joensuu, Finland. 2014.
[15] A. Ikeno, V. Varadarajan, S. Patil, J.H.L. Hansen, "UT-Scope: Speech under Lombard effect and cognitive stress," *in Aerospace Conference, 2007 IEEE*, 2007, pp. 1–7.
[16] P. Kenny, G. Boulianne, P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.
[17] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, R. Dehak, "Language recognition via i-vectors and dimensionality reduction." in *Proc. INTERSPEECH*. Citeseer, 2011, pp. 857–860.
[18] K. Wooil, J.H.L. Hansen. "Robust angry speech detection employing a TEO-based discriminative classifier combination." In *Proc. INTERSPEECH*, 2009, pp. 2019-2022.
[19] T. Hasan, R. Saeidi, J.H.L. Hansen, D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," *IEEE. ICASSP*, 2013, pp. 7663–7667.
[20] C. M. Bishop et al., *Pattern recognition and machine learning*. New York: Springer, 2006.
[21] R. Singh, M. Vatsa, A. Noore, "Intelligent biometric information fusion using support vector machine," *Soft Computing in Image Processing*. Springer, 2007, pp. 325–349.
[22] M. Vatsa, R. Singh, A. Noore, "Integrating image quality in 2*v*-svm biometric match score fusion," *International Journal of Neural Systems*, vol. 17, no. 05, pp. 343–351, 2007.
[23] Y. Freund, R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.