Automatic Recognition of Speaker Physical Load using Posterior Probability Based Features from Acoustic and Phonetic Tokens

Ming Li[12]

[1]SYSU-CMU Joint Institute of Engineering, Sun Yat-Sen University, Guangzhou, China
[2]SYSU-CMU Shunde International Joint Research Institute, Shunde, China

liming46@mail.sysu.edu.cn

Abstract

This paper presents an automatic speaker physical load recognition approach using posterior probability based features from acoustic and phonetic tokens. In this method, the tokens for calculating the posterior probability or zero-order statistics are extended from the conventional MFCC trained Gaussian Mixture Models (GMM) components to parallel phonetic phonemes and tandem feature trained GMM components. Phoneme recognizers from five different languages are employed to extract the phoneme posterior probabilities. We show that these histogram style features at both the acoustic and phonetic levels are effective and complementary for capturing the speaker physical load information from short utterances. Support vector machine is adopted as the supervised classifier. By combining the proposed methods with the OpenSMILE baseline which covers the acoustic and prosodic information further improves the final performance. The proposed fusion system achieves 70.18% and 72.81% unweighted accuracy on the validation and test set of the Munich Bio-voice Corpus for the binary physical load level recognition task in the INTERSPEECH 2014 Computational Paralinguistics Challenge.

Index Terms: physical load sub-challenge, speaker physical load recognition, posterior probability features

1. Introduction

Automatic recognition of paralinguistic information, such as speaker identity, gender, age range, emotional state, intoxication state, pathology state and cognitive load state [1, 2, 3, 4, 5] can guide human computer interaction systems to automatically understand and adapt to different user needs. Likewise such meta-information can serve as an important analytic in human decision making. For instance, the emerging broad area of behavioral signal processing aims to create quantitative characterization of typical, atypical, and distressed human behavior states across a variety of application domains including in education and health care [6, 7, 8].

Heart rate is considered as an important vital sign and feature for mobile health [9] and physical load recognition applications [10]. Although heart rate estimation from conversional Electrocardiography (ECG) signal has achieved high accuracy, it still remains a challenge and unsolved problem for non-invasive and non-contact audio-visual signals [11]. Recently, [12] and [13] show that heart rate, breathing rate and heart rate variability can be accurately determined by a laptop's or mobile phone's built-in video camera. Heart rate estimation from speech [11, 14] also starts to attract more attention since in some applications, such as the emergency call center, only speech signals are available.

For speech-based heart rate recognition, Orlikoff and Baken [15] studied the connection between speech and heart rate in 1989 and found that cardiovascular system influences the vocal fundamental frequency (F_0) when pronouncing sustained vowels based on the signal-averaging and autocorrelation analysis. Furthermore, Schuller, et.al, [11, 14] applied the openSMILE toolkit [16] to perform extraction of utterance level acoustic and prosodic features and Support Vector Machine (SVM) for the subsequent classification on the spontaneous short reading utterances. However, to our best knowledge, phonetic information has not been used in this task. Therefore, it is in this context that we explore a set of posterior probability based histogram like features from both the acoustic and phonetic tokens for the speech based heart rate recognition.

Despite the openSMILE features, several Gaussian Mixture Model (GMM) based supervectors have been proposed as features for paralinguistic speaker states recognition [2, 4, 5, 17]. These supervectors originally were proposed for speaker verification and language identification tasks but also performed well in the paralinguistic challenges. However, when the duration of speech utterance is very short (e.g. less than 2 seconds), the performance of those supervectors replying on the first-order Baum-Welch statistics (mean supervector, i-vector, Maximum Likelihood Linear Regression (MLLR) supervector, etc. [2]) drops as there are not enough feature frames to calculate the sufficient statistics. The zero-order statistics based posterior probability feature achieves better performance in these short duration scenarios with limited training data [2].

Since the utterances of the Munich Bio-voice Corpus (MBC) (the official database for the physical load sub-challenge) are indeed very short (average 1.28 seconds long), we adopt the posterior probability based histogram style features but extend the tokens from the acoustic MFCC trained GMM components to the phonetic phonemes and tandem feature trained GMM components. Furthermore, since the number of phonemes is smaller than the number of GMM components and the acoustic model used in the phoneme recognition is not trained with data from different heart rate conditions, we converted the phoneme posterior probabilities into tandem features [18, 19] and then apply GMM on top of it to generate tokens using the MBC training data. This is also motivated by the hierarchical phoneme posterior probability estimator in [20]. In this setup, the entire GMM statistics calculation remains the same except that the GMM model is trained on the tandem features.

This phoneme posterior probability (PPP) based tandem feature has been reported to be effective in speech recognition

This research is funded in part by CMU-SYSU Collaborative Innovation Research Center and the SYSU-CMU Shunde International Joint Research Institute.

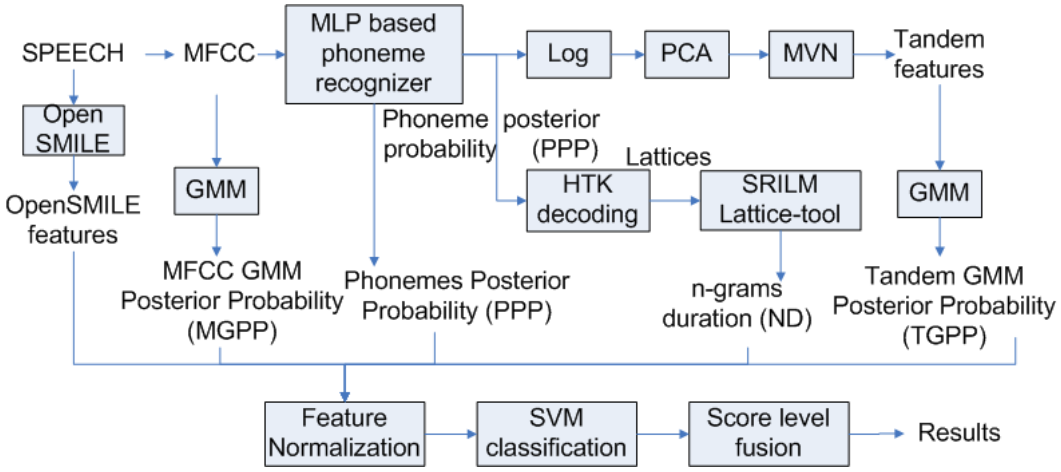

Figure 1: The system overview

[18, 19, 21], speaker verification [22] and language identification tasks[22, 23]. Here, it is used to present the phonetic information about different physical load levels. We also adopt the duration features on different 3-grams calculated from the decoded lattices to measure the duration information. Score level fusion is employed to combine multiple complementary systems together to further improve the overall system performance.

The remainder of the paper is organized as follows. The corpus and the proposed algorithms are explained in Sections 2 and 3, respectively. Experimental results and discussions are presented in Section 4 while conclusions are provided in Section 5.

2. Corpus

The database used to evaluate the proposed methods is the Munich Bio-voice Corpus (MBC) database [11, 24, 14]. The task is to classify the speaker's binary physical load level in terms of heart rate per minute (BPM) which is defined as follows: High ($\geq$90 BPM) and Low ($<$90 BPM). The mean and standard deviation of speech duration per data sample in the training and validation sets of the MBC database are 1.30 ± 0.44 s and 1.31 ± 0.39 s, respectively. Thus it is indeed a short duration database. The training data set of the MBC database (6 speakers, 385 utterances) was used for model training while the validation data set from the MBC database (6 speakers, 384 utterances) was used as the evaluation set of each subsystem as well as the fusion system in this paper. Finally, the testing data set from the MBC database (319 utterances) was evaluated. The details about the MBC database and the evaluation protocol are provided in [11, 24, 14].

3. Methods

The overview of the proposed system is demonstrated in Fig. 1. We can see that there are five different features, namely OpenSMILE feature, MFCC-GMM posterior probability (MGPP) feature, phoneme posterior probability (PPP) feature, n-gram duration (ND) feature and tandem-GMM posterior probability (TGPP) feature, followed by the same feature normalization, SVM classification and score level fusion pipeline. We first present the proposed features in section 3.1. Then section 3.2 describes the supervised classification and score level fusion methods, respectively.

3.1. Features

3.1.1. The OpenSMILE feature

The utterance level 6373 dimensional OpenSMILE feature was extracted by the OpenSMILE toolkit and provided by the 2014 Paralinguistic Challenge organizers. The details of the feature extraction is presented in [14]. Since various kinds of features, such as MFCC, loudness, auditory spectrum, voicing probability, F0, F0 envelop, jitter, and shimmer, etc., are included, this feature set can capture physical load information at both the acoustic and prosodic levels.

3.1.2. The MFCC-GMM posterior probability (MGPP) feature

For each utterance in the training and validation sets, MGPP feature extraction is performed using the Universal Background Model (UBM). Given a frame-based MFCC feature $\boldsymbol{x_t}$ and the GMM-UBM λ with M Gaussian components (each component is defined as λ_i),

$$\lambda_i = \{w_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}, i = 1, \cdots, M, \tag{1}$$

the occupancy posterior probability is calculated as follows:

$$P(\lambda_i|\boldsymbol{x_t}) = \frac{w_i p_i(\boldsymbol{x_t}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})}{\Sigma_{j=1}^{M} w_j p_j(\boldsymbol{x_t}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})}. \tag{2}$$

This posterior probability can also be considered as the fraction of this feature $\boldsymbol{x_t}$ coming from the i^{th} Gaussian component which is also denoted as partial counts. The larger the posterior probability, the better this Gaussian component can be used to represent this feature vector. The MGPP supervector is defined as follows:

$$\boldsymbol{b} = [b_1, b_2, \cdots, b_M], b_i = \frac{y_i}{T} = \frac{1}{T}\Sigma_{t=1}^{T} P(\lambda_i|\boldsymbol{x_t}) \tag{3}$$

$$MGPP_{feature} = \sqrt{\boldsymbol{b}} \tag{4}$$

Equation (3) is for calculating the zero-order Baum-Welch statistics and is exactly the same as the weight updating equation in the expectation-maximization (EM) algorithm in GMM training. In order to apply Bhattacharyya probability product (BPP) kernel [25], we adopt $\sqrt{\boldsymbol{b}}$ as our MGPP features [2].

3.1.3. The phoneme posterior probability (PPP) feature

In this PPP feature extraction, the tokens for calculating the zero-order statistics are extended from the acoustic MFCC trained GMM components to the phonetic phonemes. PPP feature is also calculated by equation (4) except that $P(\lambda_i|\boldsymbol{x_t})$ now is the posterior probability of feature $\boldsymbol{x_t}$ on the i^{th} phoneme. We believe this histogram style feature can capture the phoneme confidence and duration information for distinguishing different physical load levels. We employed the multilayer perceptron (MLP) based phoneme recognizer [26] with acoustic models from five different languages, namely Czech, Hungarian, Russian, English and Mandarin. The models for the first three languages were trained on SpeechDat-E databases and provided in [26]. Additionally, we trained the English and Mandarin based models both with 1000 neurons in all nets using the switchboard, fisher databases and the call friend, call home databases, respectively.

3.1.4. The tandem-GMM posterior probability (TGPP) feature

Since the number of phonemes is smaller than the number of GMM components and the acoustic model used in the phoneme recognition is not trained with data from different physical load conditions, we converted the PPP into tandem features [18, 19] and then apply GMM on top of it to generate tokens using the MBC training data. This is also motivated by the hierarchical phoneme posterior probability estimator in [20].

In this setup, the original PPP are converted into tandem features by log transform, principal component analysis (PCA) and mean variance normalization (MVN) [18, 19, 23] as shown in fig. 1. Then we directly consider this tandem feature as $\mathbf{x_t}$ in (2) and extract the TGPP features using (4).

3.1.5. The n-gram duration (ND) feature

Previously, the histogram stype PPP and TGPP features can be used to capture the relative occupancy and duration information for the phonemes. In this work, we also extend the tokens from phonemes to the trigrams in order to capture the context information for physical load recognition.

As shown in fig. 1, HTK toolkit [27] is used to decode the PPP features and output a lattice file for each utterance which is further processed into n-gram counts and n-gram indexes by the lattice-tool toolkit [28]. Rather than the PPP at the frame level, each trigram can span multiple frames. Therefore, in the n-gram duration (ND) feature extraction, the trigram posterior count is weighted by the trigram duration.

3.2. Classification and fusion

LIBLINEAR [29] was adopt for the SVM classification and we applied the max/min normalization (range -1 to +1) for each feature dimension on training, validation and test partitions with parameters computed only from the training partition.

Due to the limited amount of training data, we simply employed the weighted summation fusion approach with parameters tuned by cross validation. When the evaluation was performed on the testing set of the MBC database, both the training and validation sets were used for modeling and the weight vector was exactly the same as the one tuned on the validation set. It is worth noting that other advanced score fusion approaches, like the logistic regression method in the popular FoCal toolkit [30], can also be adopted here to increase the performance which is a topic for our future work.

Table 1: Performance on the validation set with different features for SVM classification and score level fusion. (The size of GMM in both the MGPP and TGPP feature extraction is 256.)

System	Features	parameter C	WA (%)	UA (%)
1	OpenSMILE	0.02	**67.45**	**67.15**
2	MGPP	0.02	61.46	61.18
3	PPP 5 languages	0.02	63.02	62.71
4	TGPP 5 languages	0.01	65.63	65.56
5	ND English	0.008	58.86	59.11
6	Fusion 2+3		65.36	65.04
7	Fusion 2+3+4		68.23	67.92
8	Fusion 2+3+4+5		68.49	68.17
9	Fusion 1+2+3+4		**70.57**	**70.18**

Table 2: Performance on the validation set using PPP and TGPP features calculated with tokens from different languages

features& Languages	PPP		TGPP	
	WA(%)	UA(%)	WA(%)	UA(%)
English	53.65	53.15	59.90	59.73
Mandarin	57.81	57.36	58.33	58.32
Czech	59.63	59.04	61.46	61.65
Hungarian	54.95	54.59	62.24	61.97
Russian	57.03	56.58	63.54	63.49
5 languages	63.02	62.71	65.63	65.56

Table 3: Confusion matrix for the binary physical load recognition on the validation set: (left) System 1, (right) System 9

	LOW	HIGH
LOW	150	49
HIGH	76	109

	LOW	HIGH
LOW	161	38
HIGH	75	110

Table 4: Confusion matrix and system performance for the binary physical load recognition of System 9 on the test set

	LOW	HIGH
LOW	116	49
HIGH	38	116

	Performance
WA(%)	72.727
UA(%)	72.814

4. Experimental results

The performance results on the validation set with different features for SVM classification are shown in Table 1. The performance is measured by weighted accuracy (WA) and unweighted accuracy (UA), respectively. First, we can see that the 6373 dimensional OpenSMILE baseline feature outperformed the 256 dimensional MGPP feature which might be because the OpenSMILE feature covers both the acoustic and prosodic information. Second, features based on the phonetic tokens (PPP and TGPP) achieved better performance compared to the acoustic tokens (MGPP). The underlying reason might be the usage of parallel phoneme recognizers from multiple languages. From Table 2, we can see that although PPP and TGPP features are effective for the physical load recognition, single language based feature or system generated lower accuracy compared to the GMPP feature and the OpenSMILE baseline. However, combing multiple systems using phonemes from different languages improved the results. Third, by comparing the results of system 3 and 4, we can find that additional stage of GMM clustering is necessary and can adapt the PPP features toward this MBC database. We only reported the ND feature result on the English recognizer (due to the lack of German acoustic model) and the results are not significantly better than the phoneme based ones

which might be because the utterance duration is too short and the feature vector becomes sparse.

Finally, by fusing the first four systems in Table 1 together, the proposed approach achieved 70.18% and 72.81% unweighted accuracy on the validation and test set, respectively. From the confusion matrix in Table 4, we can see that the proposed methods help to reduce the miss rate of the "LOW" class. In Table 4, we can see that the proposed fusion system achieved a balanced performance for both classes.

5. Conclusions

This paper presents an automatic speaker physical load recognition approach using posterior probability based features from both the acoustic and phonetic tokens. In this method, the tokens for calculating the posterior probability or zero-order statistics are extended from the conventional MFCC trained GMM components to the parallel phonetic phonemes and tandem feature trained GMM components. Phoneme recognizers from five different languages are employed to extract the phoneme posterior probabilities. We show that these histogram style features at both the acoustic and phonetic levels are effective and complementary for capturing the speaker physical load information from short utterances. By combining the proposed methods with the OpenSMILE baseline system which covers both the acoustic and prosodic information further improves the final performance.

6. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[2] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[3] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011.

[4] D. Bone, M. Li, M. P. Black, and S. S. Narayanan, "Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and gmm supervectors," *Computer speech & language*, vol. 28, no. 2, pp. 375–391, 2014.

[5] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Computer Speech & Language*, 2014.

[6] 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.

[7] M. Black, A. Katsamanis, C. Lee, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Automatic Classification of Married Couples' Behavior Using Audio Features," in *Proc. INTERSPEECH*, 2010, pp. 2030–2033.

[8] C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of Prosodic Entrainment in Affective Spontaneous Spoken Interactions of Married Couples," in *Proc. INTERSPEECH*, 2010, pp. 793–796.

[9] U. Mitra, B. A. Emken, S. Lee, M. Li, V. Rozgic, G. Thatte, H. Vathsangam, D. Zois, M. Annavaram, S. Narayanan, *et al.*, "Knowme: A case study in wireless body area sensor network design," *IEEE Communications Magazine*, vol. 50, no. 5, pp. 116–125, 2012.

[10] M. Li, V. Rozgic, G. Thatte, S. Lee, B. Emken, M. Annavaram, U. Mitra, D. Spruijt-Metz, and S. Narayanan, "Multimodal physical activity recognition by fusing temporal and cepstral information," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, no. 4, pp. 369–380, 2010.

[11] B. Schuller, F. Friedmann, and F. Eyben, "Automatic recognition of physiological parameters in the human voice: Heart rate and skin conductance," in *Proc. ICASSP*, 2013, pp. 7219–7223.

[12] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[13] C. Scully, J. Lee, J. Meyer, A. M. Gorbach, D. Granquist-Fraser, Y. Mendelson, and K. H. Chon, "Physiological parameter monitoring from optical recordings with a mobile phone," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 2, pp. 303–306, 2012.

[14] B. Schuller, S. Steidl, A. Batliner, F. Epps, J.and Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proc. INTERSPEECH*, 2014.

[15] R. F. Orlikoff and R. Baken, "The effect of the heartbeat on vocal fundamental frequency perturbation," *Journal of speech and hearing research*, vol. 32, no. 3, p. 576, 1989.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[17] M. Li, A. Metallinou, D. Bone, and S. Narayanan, "Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling," in *Proc. ICASSP*, 2012, pp. 1937–1940.

[18] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. ICASSP*, vol. 3, 2000, pp. 1635–1638.

[19] D. P. Ellis, R. Singh, and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," in *Proc. ICASSP*, vol. 1, 2001, pp. 517–520.

[20] J. Pinto, S. Garimella, M. Magimai-Doss, H. Hermansky, and H. Bourlard, "Analysis of mlp-based hierarchical phoneme posterior probability estimator," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 225–241, 2011.

[21] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, "Using mlp features in sris conversational speech recognition system," in *Proc. INTERSPEECH*, 2005.

[22] M. Li and W. Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," in *submitted to INTERSPEECH*, 2014.

[23] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "Shifted-delta mlp features for spoken language recognition," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 15–18, 2013.

[24] B. Schuller, F. Friedmann, and F. Eyben, "The munich biovoice corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production," in *Proc. Language Resources and Evaluation Conference*, 2014.

[25] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *The Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.

[26] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme," in *Proc. ICASSP*, 2006, pp. 325–328, software available at http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

[27] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*. Entropic Cambridge Research Laboratory Cambridge, 1997, vol. 2.

[28] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit," in *Proc. INTERSPEECH*, 2002.

[29] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[30] N. Brümmer, "Focal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition scoresł-tutorial and user manual," 2007, software available at http://sites.google.com/site/nikobrummer/focalmulticlass.