On the Use of Bhattacharyya based GMM Distance and Neural Net Features for Identification of Cognitive Load Levels

Tin Lay Nwe, Nguyen Trung Hieu, Bin Ma

Human Language Technology Department, Institute for Infocomm Research (I2R), A*STAR, Singapore 138632

{tlnma,thnguyen,mabin}@i2r.a-star.edu.sg

Abstract

This paper presents a method for detecting cognitive load levels from speech. When speech is modulated by different levels of cognitive load, acoustic characteristics of speech change. In this paper, we measure acoustic distance of a stressed utterance from the baseline stress free speech using GMM-SVM kernel with Bhattacharyya based GMM distance. In addition, it is believed that airflow structure of speech production is nonlinear. This motivates us to investigate better techniques to capture nonlinear characteristic of stress information in acoustic features. Inspired by the recent success of neural networks for representation learning, we employ a single hidden layer feed forward network with non-linear activation to extract the feature vectors. Furthermore, people have different reactions to a particular task load. This inter-speaker difference in stress responses presents a major challenge for stress level detection. We use a bootstrapped training process to learn the stress response of a particular speaker. We perform experiments using data sets from Cognitive Load with Speech and EGG (CLSE) provided for the Cognitive Load Sub-Challenge of the INTERSPEECH 2014 Computational Paralinguistics Challenge. The results show that the system with our proposed strategies performs well on validation and test sets.

Index Terms: cognitive load, GMM-supervector, neural net features

1. Introduction

Research on detecting task load stress is an important research topic in the area of Human Computer Interaction (HCI). Cognitive Load Sub-Challenge of the INTERSPEECH 2014 Computational Paralinguistics Challenge [1] focuses on the automatic recognition of Cognitive Load (CL) in speech. The work described in this paper aims to contribute to this challenge.

There are two main modules in cognitive load classification system. The first module is front-end feature extraction and the second module is back-end classifier. Many of the recent studies [2], [3], [4], [5] use Gaussian Mixture Model (GMM) as classifier for CL classification. Recently, Support Vector Machine (SVM) becomes popular in emotion classification [6], [7] and speaker identification [8] tasks. SVM is a novel type of learning machine, which is an approximate implementation of the method of structural risk minimization. SVM has shown to provide a better generalization performance in solving various classification problems than traditional techniques [9]. In this paper, we use SVM as the back-end classifier.

As for front-end feature extraction, much work has be done to extract reliable acoustic features for CL classification. In [2], Mel-Frequency Cepstral Coefficients (MFCC), pitch and intensity are used as features for cognitive load classification. In [3], statistics of pitch, formant, spectral slope, duration, spectral center of gravity and spectral energy spread are used as features to detect cognitive load level of drivers while they are controlling a vehicle. The study [4] investigates the effects of cognitive load on glottal parameters (open quotient, normalized amplitude quotient and speed quotient), and uses these parameters as features for cognitive load classification. The study in [10] mentions that distribution of Cognitive Load (CL) information across the bandwidth of speech has indicated that there is a large variation in the amount of CL information contained in different frequency bands of the speech signal, with the most discriminative spectral region being 0-1 kHz. Similarly, noise power is also usually unequally distributed. Based on the findings of the study in [10], the authors in [11] utilize speech features computed in disjoint bands of the spectrum and investigate the effectiveness of the multi-band approach for classifying cognitive load from speech in the presence of noise. The authors in [5], investigate the use of spectral centroid frequency (SCF) and spectral centroid amplitude (SCA) features to apply them to the problem of automatic cognitive load classification. This study shows that the spectral centroid features perform better than MFCC, pitch and intensity features.

The above studies investigate quite a number of features to discriminate between different cognitive load levels. However, the accuracies obtained by these systems are still not high enough to allow for their use outside of laboratory environments. One reason for this might be the imperfect acoustic description of speech provided by acoustic features investigated so far.

Acoustic characteristics of stressed speech is different from that of stress free neutral speech. Furthermore, there are dissimilarities in acoustic characteristics between utterances under high and low task loads. Hence, Acoustic Dissimilarity or Acoustic Distance (AD) measure of stress speech from neutral speech is useful for CL classification. In our earlier work [12], we investigate acoustic feature that characteristics AD for emotion classification. The feature characterizes AD measure between an emotion utterance and neutral speech. AD measure is formulated using Bhattacharyya distance based GMM-supervectors [8]. In this work, we employ this Bhattacharyya distance based AD measure for classifying cognitive load.

Teager [13] suggests that the true source of sound production is vortex-flow interactions which are nonlinear. It is believed that changes in vocal system physiology induced by stressful conditions such as muscle tension will affect the vortex-flow interaction patterns in the vocal tract [14]. Therefore, nonlinear speech features are important to classify different levels of task load stress. We investigate better techniques to

capture nonlinear characteristic of stress information in acoustic features. Inspired by the recent success of neural networks for representation learning, we employ a single hidden layer feed forward network with non-linear activation to extract the feature vectors.

Different people have different reactions to a particular task load. A stress model will be more accurate if it can learn the characteristics of stress response from test utterance. We use a bootstrapped training process to learn the stress response of test utterances.

In this paper, we explore features which are shown to be complementary to the baseline features provided by organizers of INTERSPEECH 2014 Computational Paralinguistics Challenge [1]. We propose to use features that characterize AD measure based on Bhattacharyya distance. Furthermore, we extract feature vectors using a single hidden layer feed forward network with non-linear activation to learn non-linear acoustic characteristics of stress utterances. We use the classifier with bootstrapping strategy [15] to learn characteristics of stress response from test samples.

The rest of the paper is organized as follows. Section 2 presents Bhattacharyya distance based GMM supervectors that characterize AD measure. Section 3 explain feature extraction using neural network. Section 4 presents our classification process with bootstrapping process. Section 5 describes database and baseline features. Section 6 presents experimental results and Section 7 concludes the paper.

2. GMM-supervector

The GMM-supervector can be considered of as a mapping between an utterance and a high-dimensional vector through a kernel [16]. Kernels are important components for SVM learning. It is a method of using a linear classifier to solve a non-linear problem by nonlinearly mapping the original observations into a higher-dimensional space, where a linear classifier is subsequently used. This makes linear classification in the new feature [17] set.

2.1. Gaussian mixture model (GMM)

Gaussian mixture model (GMM) is the most effective way to model the spectral distribution of speech. A GMM-supervector characterizes speaker's stress information by the GMM parameters such as the mean vectors, covariance matrices and mixture weights. The density function of a GMM is defined as in equation (1).

$$p\left(x\right) = \sum_{i=1}^{M} \omega_i f\left(x|m_i, \Sigma_i\right) \tag{1}$$

where $f\left(.\right)$ denotes the Gaussian density function. And, m_i, Σ_i and ω_i are the mean, covariance matrix and weight of i^{th} Gaussian component, respectively. M is number of Gaussian mixtures. And, x is a D-dimensional acoustic feature vector. GMM-supervector formulation using GMM-SVM kernels based Bhattacharyya based GMM distance is as follows.

2.2. GMM-supervector with Bhattacharyya based kernel

Bhattacharyya distance is a separability measure between two Gaussian distributions [18]. The Bhattacharyya distance between the two probability distributions is defined as in equation (2) [8].

In equation (2) Σ_i^a and Σ_i^b are the adapted covariance matrices. And, m_a^i and m_b^i are the adapted mean vectors. Σ_i^u is the covariance matrix of the Universal Background Model (UBM). p_a and p_b are the probabilistic models, GMM_a and GMM_b, respectively.

The first term of equation (2) gives the class separability due to the difference between class means, while the second term gives the class separability due to the variance between class covariance. Based on the first two terms, Bhattacharyya distance based kernel is formulated as in equation (3) [8]. Based on this kernel, the i^{th} subvector of the GMM-supervector is formulated as in equation (4)[8]. GMM-supervector with Bhattacharyya based kernel is obtained by stacking all i^{th} subvectors of equation (4).

$$g^{Bhat}(m_i, \Sigma_i) = \begin{bmatrix} \left(\frac{\Sigma_i^\lambda + \Sigma_i^u}{2}\right)^{-1/2} \left(m_i^\lambda - m_i^u\right) \\ \\ \text{diag}\left(\left(\frac{\Sigma_i^\lambda + \Sigma_i^u}{2}\right)^{1/2} \left(\Sigma_i^\lambda\right)^{-1/2}\right) \end{bmatrix} \tag{4}$$

If we look at equation (4), the first term reflects the dissimilarity between mean of a stress utterance and that of a UBM. This mean statistical dissimilarity gives the major characteristics of the probabilistic distance. And, this term represents the distance of a stress utterance from a reference neutral UBM. If a reference UBM is trained using stress free neutral utterances, this term is to measure the distance of a stress utterance from a baseline stress free UBM. Besides the first-order statistics of mean, the second-order statistics of covariance matrices describing the shape of the spectral distribution is also useful to measure the distance. If we look at the second term of equation (4), it represents the ratio between covariance of a UBM and that of a stress utterance. In other words, this second term describes dissimilarity measure in terms of spectral shape.

3. Neural net features

The baseline features extracted with the openSMILE toolkit [19] and the GMM-supervector introduced in Section 2 are often considered to be low-level features as they represent various aspects of speech including phonetic contents, speaker identities and emotions etc. The features are not specifically formulated to capture the cognitive load pattern in speech. We thus believe that higher abstract features which are explicitly designed to discriminate various cognitive load patterns will bring further improvements to the identification system. In this section, a feed forward neural network with a single hidden layer is proposed, which is then used to transform the low-level features to more discriminative features. The network architecture is illustrated in Figure 1, where $\tilde{\mathbf{x}}$ is the corrupted version of the input feature vector $\mathbf{x}$, $\mathbf{h}$ is the hidden layer with the number of nodes equals to the dimension of $\mathbf{x}$ (over complete), and $\mathbf{y}$ is the output layer with 3 nodes corresponding to the cognitive load levels.

Formally defined:

$$\tilde{x}_i = \begin{cases} x_i & \text{with probability 0.5;} \\ 0 & \text{with probability 0.5} \end{cases} \tag{5}$$

$$\mathbf{h} = sigmoid\left(\mathbf{W}^{(1)}\tilde{\mathbf{x}} + \mathbf{b}^{(1)}\right) \tag{6}$$

$$\mathbf{y} = softmax\left(\mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}\right) \tag{7}$$

$$\Psi^{Bhat}\left(p_a \| p_b\right) \approx \frac{1}{8}\sum_{i=1}^{M}\left\{\left[\left(\frac{\Sigma_i^a+\Sigma_i^u}{2}\right)^{-1/2}\left(m_i^a-m_i^u\right)\right]^T\left[\left(\frac{\Sigma_i^b+\Sigma_i^u}{2}\right)^{-1/2}\left(m_i^b-m_i^u\right)\right]\right\}$$
$$+\frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\left(\frac{\Sigma_i^a+\Sigma_i^u}{2}\right)^{1/2}\left(\Sigma_i^a\right)^{-1/2}\left(\frac{\Sigma_i^b+\Sigma_i^u}{2}\right)^{1/2}\left(\Sigma_i^b\right)^{-1/2}\right]$$
$$+\frac{1}{2}\sum_{i=1}^{M}\left\{\frac{\omega_i^u}{\omega_i^a}\frac{\omega_i^u}{\omega_i^b}\right\}-\sum_{i=1}^{M}\ln\left\{\omega_i^u\right\}-M \tag{2}$$

$$K^{Bhat}\left(X_a, X_b\right) = \sum_{i=1}^{M}\left\{\left[\frac{1}{2}\left(\frac{\Sigma_i^a+\Sigma_i^u}{2}\right)^{-1/2}\left(m_i^a-m_i^u\right)\right]^T\left[\frac{1}{2}\left(\frac{\Sigma_i^b+\Sigma_i^u}{2}\right)^{-1/2}\left(m_i^b-m_i^u\right)\right]\right\}$$
$$+\sum_{i=1}^{M}\mathrm{tr}\left[\left(\frac{\Sigma_i^a+\Sigma_i^u}{2}\right)^{1/2}\left(\Sigma_i^a\right)^{-1/2}\left(\frac{\Sigma_i^b+\Sigma_i^u}{2}\right)^{1/2}\left(\Sigma_i^b\right)^{-1/2}\right] \tag{3}$$

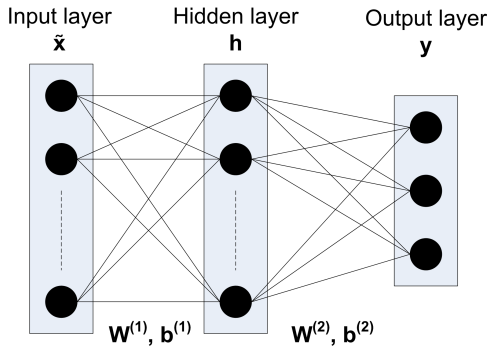

Figure 1: *Neural network architecture*

where $\tilde{x}_i$ and x_i are respectively the i^{th} element of $\tilde{\mathbf{x}}$ and $\mathbf{x}$, $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are the weight matrices, $\mathbf{b}^{(1)}$ and $\mathbf{b}^{(2)}$ are the bias vectors. The network parameters are first initialized with the stacked denoise auto-encoder approach, then they are iteratively fine-tuned using the stochastic gradient descent algorithm until the errors on the validation set stop decreasing. Once the network is trained properly, for each input feature vector $\mathbf{x}$, the corresponding $\mathbf{h}$ is then used as the new feature vector.

4. Classification with bootstrapping process

To allow for a fair comparison with the baseline system we stick to the classification scheme suggested by the challenge organizers. That is we also use the WEKA data mining toolkit [20] and employ a linear kernel Support Vector Machines (SVM) with Sequential Minimal Optimisation (SMO). The following is our bootstrapping process [15] to learn characteristics of the stress response from test samples.

Firstly, we train SVM models using training data. These models learn characteristics of the stress response from training samples. We use these models to classify the CL levels of utterances in test set. Then, we select the 20% of the test samples with the highest scores to the models for bootstrapped training. We combine training samples and 20% of bootstrapped samples to re-train the models referred to as bootstrapped models. Finally, we perform classification on test set using bootstrapped models.

5. Database and baseline features

The Cognitive Load Sub-Challenge uses Cognitive Load with Speech and EGG (CSLE) database [21] for the detection of 3 different cognitive load levels: low(L1), medium(L2) and high(L3). It consists of 2418 utterances. Average length of the utterances is 4 seconds. The challenge provides 3 datasets: training, validation and test sets. Labels of the CL for training and validation sets are available to participants. However, labels for test set are unknown to the participants.

From each utterance, a total of 6373 static features functionals of low-level descriptor (LLD) contours are extracted. TUMs open-source openSMILE feature extractor [19] in its recent 2.0 release [22] is used to extract features. Organizers provide these feature sets for all 3 data sets. The feature sets are referred to as baseline feature sets. More details about the database and the baseline features can be found in [1].

6. Experiments and results

We conduct several experiments to investigate the effectiveness of proposed features. Firstly, we examine the capabilities of individual features sets: Bhattacharyya based GMM supervectors (Bhat-GMM-Sup) and Neural Net (NN) features to classify CL. Then, we integrate our proposed feature sets to the baseline feature set and observe the improvement on classification accuracy. Finally, we investigate the CL classification performance of our bootstrapped classifier. In all experiments, we use Unweighted Average Recall (UAR) [1] to measure the classification accuracies.

Each utterance is divided into 20ms frames with 10ms overlapping. Each frame is multiplied by a Hamming window to minimize signal discontinuities at the end of each frame. From each frame, we extract Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficient (LPCC) and Perceptual Linear Prediction Coefficients (PLPC) features. Each feature has 12 coefficients and their first derivatives. We form a feature vector for each frame by concatenating all three MFCC, LPCC and PLPC features. As each feature has a total of

24 coefficients, a feature vector of a frame has 72 coefficients.

In all experiments, we use open-source classifier implementations of Support Vector Machines (SVM) from the WEKA data mining toolkit [20] as classifier. Linear kernel SVM is used, as it is known to be robust against over-fitting. For simplicity we keep the complexity parameter C fixed at 0.008. We presents the accuracy of the baseline system with our classifier setting in 8^{th} row of Table 1. We find that performance of our baseline system is 0.85% lower than that of the baseline system of the organizers [1]. The reason is the difference in complexity parameter C in SVM classifier.

To formulate GMM-supervectors, we extract the MFCC, LPCC and PLPC features mentioned above from each utterance. Then, we use maximum a posteriori (MAP) criterion [23] to adapt a GMM model from a Universal Background Model(UBM) for each utterance. We train UBM via EM algorithm [24]. We use Rich Transcription 2007 (RT-07) Evaluation dataset to train UBMs [25]. This RT-07 dataset includes neutral speech utterances recorded during a meeting. We use this RT-07 set to train neutral UBMs to measure AD. We adapt the mean and covariance only. Once we have an adapted GMM model, we formulate GMM-supervectors using the techniques mentioned in Section 2. CL classification accuracies on validation dataset of Cognitive Load Sub-Challenge using Bhattacharyya based GMM supervectors (Bhat-GMM-Sup) are presented in Table 1. We present the results for different Gaussian mixtures. We use training set to train the CL models.

Table 1: *Average accuracies (%) using individual feature sets on development set*

Setting	UAR[%]
Bhat-GMM-Sup (4 mix.)	40.25
Bhat-GMM-Sup (8 mix.)	42.82
Bhat-GMM-Sup (16 mix.)	52
Bhat-GMM-Sup (32 mix.)	53.69
Bhat-GMM-Sup (64 mix.)	53.65
NN-features	65.16
baseline features	62.35

Results show that classification performance improves when we increase the number of mixtures for GMM. We find that these individual feature sets can not perform better than baseline features. The reason is that baseline feature is combination of many types of acoustic features. And, each type has its own capability to characterize stress information of an utterance. AD measure using Bhattacharyya based GMM supervectors can be taken as a particular type of acoustic feature that characterize stress information in terms of AD measure from a neutral reference model.

Next, we extract features from neural network. Input of neural network is combination of baseline features and Bhattacharyya based GMM supervectors(Bhat-GMM-Sup). We select the Bhat-GMM-Sup with 16 mixtures for combination. We extract over complete feature set. Performance of Neural Net (NN) feature is presented in 7^{th} row of Table 1. The result shows that NN features perform better than the baseline system.

We observe the contribution of complementary stress information by our features by integrating them to the baseline feature. First, we integrate Bhat-GMM-Sup feature to baseline features and perform CL classification. The result is shown in 4^{th} row of Table 2. The result shows that UAR improvement on more than 2.5% absolute over baseline system which is shown in 3^{rd} row of Table 2. We can see that Bhat-GMM-Sup feature provides additional stress information in terms of AD measure. Next, we integrate Neural Net (NN) features. The result on 5^{th} row of Table 2 shows that UAR further improves 1% absolute.

We observe the performance of Bootstrapping (BS) method. We use Random Forest classifier from the WEKA data mining toolkit [20] to select bootstrapped samples. Random Forest classifier has 1000 trees in total. The first 20% of the samples with the highest scores from test set are selected as bootstrapped samples. The result using Bootstrapping (BS) method is shown in 6^{th} row of Table 2. We find that performance improvement by bootstrapping process is not significant. The reason is that bootstrapping process needs reasonable numbers of bootstrapped samples for models to learn stress characteristics from test samples well. We could select more samples (example 30%) to see the improvement.

Finally, we conduct experiments on test set. CL labels of the test set is unknown to the participants of Cognitive Load Sub-Challenge. we submit result of the test set online to observe the CL classification accuracy. Out of five trials which are allowed to the participants to try, we try 2 trials and present the best result on 9^{th} row of Table 2. We find that our UAR is 0.1% lower than the UAR of the organizer's baseline system. The reason could be complexity parameter C of SVM system [1].

Table 2: *Results on classifying cognitive loads for Cognitive Load Sub-Challenge of COMPARE 2014*

Development Set	
Setting	UAR[%]
Baseline	62.35
Baseline + GMM-Sup	65.01
Baseline + GMM-Sup + NN-Features	66
Baseline + GMM-Sup + NN-Features with BS	66.05
Test Set	
Setting	UAR[%]
Baseline + GMM-Sup + NN-Features with BS	61.5

7. Conclusions

We have presented an approach to employ Acoustic Distance (AD) measure in GMM-supervector formulation for SVM classifier to classify cognitive loads. We formulate higher abstract features using neural network to investigate nonlinear acoustic characteristics of stress utterances. Finally, we employ the bootstrapped process in which stress models learn characteristics of the stress responses from test samples. We find that our proposed approaches provide complementary information when we integrate our approaches to the baseline system.

8. References

[1] Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., and Zhang, Y., "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load," To appear in Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, p. 5 pages. ISCA, Sept. 2014.

[2] Yin, B., Ruiz, N., Chen, F. and Khawaja, M. A.,"Automatic Cognitive Load Detection From Speech Features", OZCHI, volume 251 of ACM International Conference Proceeding Series, page 249-255. ACM, 2007.

[3] Boril, H., Sadjadi, O., Kleinschmidt, T., and Hansen, J. H. L., "Analysis and detection of cognitive load and frustration in drivers' speech", Proceedings of Interspeech'10, pp. 502505, Makuhari, Chiba, Japan, 2010.

[4] Yap, T. F., Epps, J., Choi, E., and Ambikairajah, E., Glottal Features For Speech-Based Cognitive Load Classification, Proc. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP10), Dallas, USA, pp. 5234-5237, March 2010.

[5] Le, P. N., Ambikairajah, E., Epps, J., Sethu, V., amd Choi, E. H. C., "Investigation Of Spectral Centroid Features For Cognitive Load Classification", Speech Communication 53 (4), 540-551, 2011.

[6] Wu, S., Falk, T., Chan, W., "Automatic Speech Emotion Recognition Using Modulation Spectral Features". Speech Communication 53(5):768-785, 2011.

[7] Schuller B., Reiter S., Mueller R., Al-Hames M., Lang M., Rigoll G. "Speaker-Independent Speech Emotion Recognition by Ensemble Classification". Proc. ICME 2005, Amsterdam, Netherlands, 2005.

[8] You, C., Lee, K.A., and Li, H., "GMM-SVM Kernel With A Bhattacharyya-Based Distance For Speaker Recognition" IEEE Trans. Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1300 - 1312, 2010.

[9] Chen, L., Mao, X., Xue, Y-L., and Cheng, L. L., "Speech Emotion Recognition: Features and Classification Models". Digital Signal Processing 22(6): 1154-1160, 2012.

[10] Le, P N., Ambikairajah, E., Choi, E., and Epps, J., A Non-Uniform Subband Approach to Speech-Based Cognitive Load Classification, in Proc. 7th International Conference on Information Communication and Signal Processing, ICICS, Macau., Dec. 2009.

[11] Le, N. P., Epps, J. R., Ambikairajah, E., and Sethu. V., "Robust Speech-Based Cognitive Load Classification Using a Multi-band Approach", The Proceedings of APSIPA ASC 2010, Asia-Pacific Signal Processing Association, Hong Kong, presented at Asia-Pacific Signal Processing Association Conf., Singapore, 14 - 17 December 2010.

[12] Nwe, T. L., Nguyen, T. H., and Limbu, D. K., "Bhattacharyya Distance Based Emotional Dissimilarity Measure For Emotion Classification" ICASSP 2013: 7512-7516.

[13] Teager, H. M., Some Observations on Oral Air Flow During Phonation, IEEE Trans. Acoustics, Speech, Signal Processing, vol. ASSP-28, no. 5, pp. 599601, Oct. 1980.

[14] Zhou, G., Hansen, J. H. L., Kaiser, J. F., "Nonlinear Feature Based Classification of Speech under Stress", IEEE Transactions on Speech & Audio Processing 9, 201-216, 2001.

[15] Tzanetakis, G., Song-specific Bootstrapping of Singing Voice Structure, IEEE International Conference On Multimedia And Expo, 2004.

[16] Campbell, W.M., Sturim, D.E., Reynolds, D.A., and Solomonoff, A., "SVM based Speaker Verification Using A GMM Supervector Kernel and NAP Variability Compensation", in Proc. of ICASSP, pp. 97-100, France, 2006.

[17] Vaiciukynas, E., Gelzinis, A., Bacauskiene, M., Verikas, A., and Vegiene, A., "Exploring Kernels in SVM-Based Classification of Larynx Pathology from Human Voice", Proceedings of the 5th International Conference on Electrical and Control Technologies ECT-2010, May 6-7, 2010, Kaunas, Lithuania

[18] Mak, B., and Barnard, E., "Phone Clustering Using The Bhattacharyya Distance". ICSLP 1996

[19] Eyben, F., Wollmer, M., and Schuller, B., openSMILE The Munich Versatile and Fast Open-Source Audio Feature Extractor, in Proc. of the 18th ACM International Conference on Multimedia, MM 2010. Florence, Italy: ACM, 2010, pp. 14591462.

[20] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I., The WEKA Data Mining Software: An Update, SIGKDD Explorations, vol. 11, 2009.

[21] Yap, T. F., "Speech Production Under Cognitive Load: Effects and Classification," Ph.D. dissertation, The University of New South Wales, Sydney, Australia, 2012.

[22] Eyben, F., Weninger, F.,Gro, F., and Schuller, B., Recent Developments in Opensmile, the Munich Open-Source Multimedia Feature Extractor, in Proc. of the 21st ACM International Conference on Multimedia, MM 2013, Barcelona, Spain, October 2013, pp. 835838.

[23] Gauvain, J. L., and Lee, C. H., "Maximum A Posteriori Estimation For Multivariate Gaussian Mixture Observations Of Markov Chains", IEEE Trans. Speech Audio Process., vol. 2, no. 2, pp. 291-298, 1994.

[24] Dempster, A.P., Laird, N.M., Rubin, D.B., "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society, Series B, 39: 1-38, 1977.

[25] Spring 2007 Rich Transcription Meeting Recognition Evaluation Plan, [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2007/docs/ rt07-meeting-eval-plan-v2.pdf