A PLLR AND MULTI-STAGE STAIRCASE REGRESSION FRAMEWORK FOR SPEECH-BASED EMOTION PREDICTION

Zhaocheng Huang^{1,2} *and Julien Epps*^{1,2}

[1]School of Electrical Engineering and Telecommunications, UNSW Australia and [2] Data61, CSIRO, Australia
zhaocheng.huang@unsw.edu.au, j.epps@unsw.edu.au

ABSTRACT

Continuous prediction of dimensional emotions (e.g. arousal and valence) has attracted increasing research interest recently. When processing emotional speech signals, phonetic features have been rarely used due to the assumption that phonetic variability is a confounding factor that degrades emotion recognition/prediction performance. In this paper, instead of eliminating phonetic variability, we investigated whether Phone Log-Likelihood Ratio (PLLR) features could be used to index arousal and valence in a pairwise low/high framework. A multi-stage staircase regression (SR) framework which enables fusion at three different stages is also investigated. Results on the RECOLA database show that PLLR outperforms EGEMAPS features for arousal and valence. Interestingly, long-term averaged PLLR proved to be more robust and emotionally informative than local frame-level PLLR, which contains more phoneme-specific information. Within the multi-stage SR framework, PLLR yielded an 8.2% and 11.6% relative improvement in CCC for arousal and valence respectively, showing great promise for including phonetic features in emotion prediction systems.

Index Terms — Phone log-likelihood ratio, staircase regression, relevance vector machine, emotion prediction

1. INTRODUCTION

Continuous prediction of emotion dimensions such as arousal and valence at frame basis has become an emerging area of research recently within the affective computing community [1]. The dimensional representation of emotions is more advantageous than categorical representation such as anger and neutral for capturing complex and subtle variations in emotional states [2]. In line with this popularity, annual audio-visual emotion challenges [3], [4], [5] were held to motivate and drive research towards more robust and effective systems, among which audio features, especially spectral and prosodic features [6], have proved critical.

However, the frame-level acoustic features adopted as baseline features for these challenges, contain phonetic variability, arising from variations in short-term features spanning different phonemes. This variability has been shown to have a negative impact on emotion recognition systems [7], [8]. Attempts to mitigate phonetic variability include the calculation of functionals (long-term statistics of short-term features) [9], lexical normalization [10], and examination of acoustic features from specific phonemes or phoneme classes [11], [12], [13], [14]. Moreover, there are multiple emotion recognition systems that explicitly model emotional phonemes, showing better performance compared with phoneme independent systems [15], [16], [17]. The majority of aforementioned studies have proceeded from an approach of segmenting speech on a per-phoneme basis and then applying machine learning. However, to the best of our knowledge, no studies have investigated direct usage of phonetic features for speech based emotion recognition.

A promising approach for prediction that was originally proposed for depression speech is known as Gaussian Staircase Regression [18]. The staircase regression (SR) approach, which allows pairwise comparison of low and high classes, has been shown effective for depression [19], [20] and more recently for emotion prediction [21]. Considered from a more general perspective, the pairwise comparison basis of SR suggests that low-high comparisons of arousal or valence could be made on a per-phoneme basis, and in this context it is of interest to consider features that contain phone-specific information, i.e. to make frame-by-frame comparisons between low and high arousal/valence on a per-phoneme basis.

In this paper, we investigate direct utilization of phonetic features, i.e. Phone Log-Likelihood Ratio (PLLR), in a proposed multi-stage SR framework for continuous emotion prediction.

2. RELATED WORK

Phonetic variability is notorious for its impact on speech emotion recognition. For instance, it has been suggested in [10] that phonetic variability of lexical content dominates acoustic features rather than speaker and emotion variability in emotional speech. Three common approaches have been adopted to remove the variability. Among the most widely used method is to calculate functionals, i.e. long-term statistics, of short-term features, thereby being less sensitive to their local variations caused by phonemes. A second approach is to explicitly compensate the variability via normalization, e.g. whitening transformation for each phoneme in [10]. A third common approach is to segment emotional speech into phonemes or phoneme classes, from which acoustic features are extracted and processed on a per-phoneme/class basis. It was found that vowels are more conducive to emotion classification in [11], [12], [13], while spectral features extracted from consonants are more effective in [14].

Apart from this, there are some studies investigating phoneme-level emotional models, exploiting the discriminative nature of some phonemes [16], [17], [22]. For example, emotion-specific vowels of Mexican Spanish speech were modelled by HMMs and the number of emotional vowels was counted [16]. Most of the abovementioned studies were generally done for classification of categorical emotions such as anger and happiness.

For dimensional emotions, in [17], an Automatic Speech Recognition (ASR) engine was used to generate a phonetic transcription, based on which phoneme-level emotion models were built for binary classification of emotion dimensions. A recent investigation found discriminative capabilities in phonetic syllables, from which acoustic features were extracted, achieving

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

state-of-the-art performance for predicting emotion dimensions using SVR at turn-level on the VAM and SEMAINE corpora [15].

Phonetic information has been suggested to be emotionally discriminative in previous studies. However, to the best of our knowledge, no studies have investigated the direct use of phonetic information for emotion prediction on a per-frame basis. Previous studies have only evaluated acoustic features within phonemes, which may be incapable of capturing the phonetic description of emotional speech [23]. We speculate that features based on phonetic information could be complementary to existing acoustic features, because they may carry emotion-related information different from that of spectral and prosodic features. Among the most popular phonetic features are the Phone Log-Likelihood Ratio (PLLR) features, which are widely used in state-of-the-art language identification systems, e.g. [24].

The Staircase Regression framework was first proposed for depression prediction in [18], where data corresponding to intervals of the rating scale were grouped into several pairs of low-high classes, and the log mean likelihood ratio (LMLR) between the low and high partition was calculated. The LMLR from each low-high class pair was then used in regression modeling, to predict depression BDI scores. Relevance Vector Machine Staircase Regression (RVM-SR), based on the same idea, was found to be effective for depression prediction [20] as well as more recently for emotion prediction [21]. Staircase methods in general have been found to effectively exploit complementary information using fusion [21], which may be helpful herein for integrating PLLR features for continuous emotion prediction.

3. SYSTEM OVERVIEW

3.1 System Overview

In this paper, only the AVEC baseline EGEMAPS features and the proposed PLLR features were considered. For both regression and classification modeling, Relevance Vector Machine (RVM) was selected due to its effectiveness for emotion prediction [25]. While Support Vector Machine/Regression yields a sparse representation of instances, RVM maintains the sparse representation for features. The proposed system (Fig. 1) enables fusion at three different levels, namely feature-level, classifier-level and score-level.

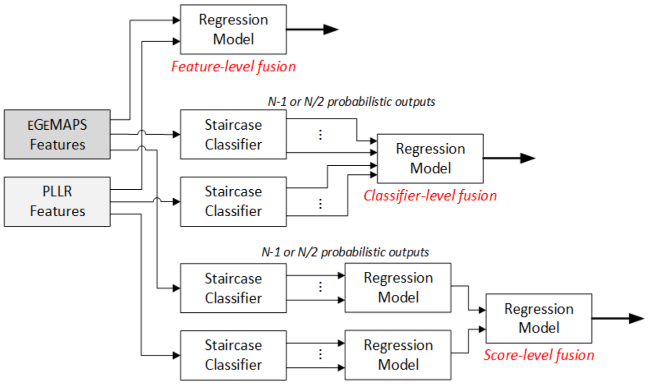


Fig. 1. *Proposed emotion prediction system (after [20])*

3.2. Phone Log-Likelihood Ratio (PLLR) Feature

Given a phone decoder with M phonemes, each of which has been modelled by one Hidden Markov Model (HMM) with S states, the posterior probability for each state s ($1 < s < S$) of each phoneme model m ($1 < m < M$) at each frame t is denoted as $p_{t,s}(m)$. Then the posterior probabilities of each phoneme are summed across all states in (1) before calculating the PLLR using (2) [24]:

$$p_t(m) = \sum_{\forall s} p_{t,s}(m) \tag{1}$$

$$PLLR_t(m) = \log \frac{p_t(m)}{\frac{1}{(M-1)}\sum_{\forall j\neq m} p_t(j)} \tag{2}$$

In (2), the numerator represents probability of phoneme m, whereas the denominator denotes the average probability of all phonemes exclusive of phoneme m. The ratio between the two provides a probabilistic measure for the presence of phoneme m. Taking the log of the ratio enables the measure to be more Gaussian-distributed and less bounded [26]. In the emotion prediction context, PLLR features (i) provide an indication of the most relevant phoneme for a given frame (allowing phoneme-specific modelling) and (ii) as a feature set provide a kind of 'positioning' of the current frame among all phonemes. Similarly, a bag-of-audio-word approach has recently been proposed to capture the 'positioning'-like information in the codebook, where the entire acoustic space of low-level descriptors is clustered [27].

3.3. Multi-Stage Staircase Regression

3.3.1. Relevance Vector Machine for Regression and Classification

A general form for RVM regression can be found in (3), where RVM searches for a weight for each feature dimension [28].

$$y(\boldsymbol{t}_*|\boldsymbol{x}_*,\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_*)+\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\sim\mathcal{N}(0,\sigma^2) \tag{3}$$

$\boldsymbol{\epsilon}$ is the trained noise and t is the predicted score. To enforce sparsity on the weights $\boldsymbol{w}$, the weights are given a prior distribution of zero-mean Gaussian, i.e. $w_k\sim\mathcal{N}(0,\alpha_k^{-1}), k\in[1,\ldots,K]$, where K is the feature dimensionality.

During the training process, RVM aims to maximize the posterior probability of all parameters given the training data.

$$p(\boldsymbol{w},\boldsymbol{\alpha},\sigma^2|\boldsymbol{t}) = p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{\alpha},\sigma^2)p(\boldsymbol{\alpha},\sigma^2|\boldsymbol{t}) \tag{4}$$

The first term on the right hand side specifies normal distributions over the weights $\boldsymbol{w}$, controlled by $\boldsymbol{\alpha},\sigma^2$. Accordingly, we maximize the posterior probability $p(\boldsymbol{\alpha},\sigma^2|\boldsymbol{t})$, which can be further reformulated as a type-II maximum likelihood problem in (5) via Bayes' rule.

$$(\boldsymbol{\alpha}_{MP},\sigma^2_{MP}) = \underset{\boldsymbol{\alpha},\sigma^2}{\operatorname{argmax}}\,\mathcal{L}(\boldsymbol{\alpha},\sigma^2) = \underset{\boldsymbol{\alpha},\sigma^2}{\operatorname{argmax}}\,p(\boldsymbol{t}|\boldsymbol{\alpha},\sigma^2) \tag{5}$$

After determining $\boldsymbol{\alpha}_{MP},\sigma^2_{MP}$, the Gaussian-distributed weights $p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{\alpha}_{MP},\sigma^2_{MP})\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)$ are sparse, because the majority of α_i tend to be infinity, i.e. zero for the corresponding weights w_i. Given test features $\boldsymbol{x}_*$, predictions become:

$$y(\boldsymbol{t}_*|\boldsymbol{x}_*,\boldsymbol{w}) = \boldsymbol{\mu}^T\boldsymbol{\phi}(\boldsymbol{x}_*) \tag{6}$$

Similarly to RVM regression, RVM classification searches for the most relevant weights for each single feature, but there is no error term $\boldsymbol{\epsilon}$ in (3), i.e. there is no parameter σ^2. Introduction of the sigmoid function to the outputs $\mathcal{S}(y) = 1/(1+e^{-y})$ leads to a Bernoulli distributed likelihood function, and therefore training of a RVM classifier involves maximizing the posterior probability $p(\boldsymbol{w}|\boldsymbol{c},\boldsymbol{\alpha})$ using Laplace's method [28]. Given test features $\boldsymbol{x}_*$, classification becomes

$$p(c_*|\boldsymbol{x}_*,\boldsymbol{w}) = \mathcal{S}\big(y(\boldsymbol{x}_*,\boldsymbol{w}_{MP})\big) = \mathcal{S}\Big(\boldsymbol{w}_{MP}^T\boldsymbol{\phi}(\boldsymbol{x}_*)\Big) \tag{7}$$

where c is the emotional class and $p(c_*|\boldsymbol{x}_*,\boldsymbol{w})$ can be regarded as a probabilistic output for binary classification.

3.3.2. Relevance Vector Machine Staircase Regression

The idea behind RVM-SR is to make pairwise comparisons between different low-high partitions, and to incorporate this information into regression modeling. Thus, it is a combination of multiple RVM classifiers and regression.

In RVM-SR, firstly emotion ratings are evenly partitioned into N partitions based on percentiles of arousal/valence scores on training data. Data corresponding to the N partitions are then grouped into low-high pairs to train different RVM classifiers C_l^h, where h denotes partitions of data from high arousal/valence, whilst l denotes low arousal/valence. For instance, $C_{l=1}^{h=2:N}$ denotes a classifier that is trained with data from the 1st partition is considered as the "low" class and data from the 2nd to Nth partitions are considered as the "high" class. The existing SR [21] therefore trains N-1 classifiers, as seen in (8).

Staircase 1: $$C_{l=1:i}^{h=i+1:N}, i \in [1,2,\dots,N-1] \tag{8}$$

Apart from (8), we proposed in this paper three additional types of staircases. The *Staircase 2* starts with comparing the most extreme cases where the data with the lowest emotion ratings are considered as "low", whereas data with the highest ratings are considered as "high". The *Staircase 3* also starts with comparing the most extreme cases, but adding more data for training as proceeds. The *Staircase 4* compares high and low classes separated by the mean of the emotion ratings for small i, but removing the extreme data partitions for larger i.

Staircase 2: $$C_{l=i}^{h=N-i+1}, i \in [1,2,\dots,N/2] \tag{9}$$

Staircase 3: $$C_{l=1:i}^{h=N-i+1:N}, i \in [1,2,\dots,N/2] \tag{10}$$

Staircase 4: $$C_{l=i:N/2-1}^{h=N/2+1:N-i+1}, i \in [1,2,\dots,N/2] \tag{11}$$

After training all classifiers, the probabilistic outputs $p(c_*|\boldsymbol{x}_*, \boldsymbol{\alpha}_{MP})$ from equation (7) were used for RVM regression modeling. Unlike the existing type of staircase, the three proposed staircases can be complementary because they contain information covering the most extreme cases and the most confusing cases.

4. EVALUATION

4.1. Database

Experiments in this paper were evaluated using the AVEC2016 database [5]. The AVEC 2016 database was selected from the *Remote Collaboration and Affective Interaction* (RECOLA) corpus [29]. It is a spontaneous multimodal corpus collected in scenarios where two French speakers complete a survival task together through a video conference. This database was chosen because it is large and has been widely used and allows comparison with other work for predicting arousal and valence. There are recordings of 5-minute length from 27 subjects, which were evenly divided into training, development and test partitions. As we did not have access to gold standard ratings from the test set data, we adopted only the training and development partition. In the database, frame-level annotations of arousal and valence are provided at every 40 milliseconds per file.

4.2. Experimental Settings

For PLLR feature extraction, we used the BUT phoneme recognizer (Hungarian) [30] to calculate 59-dimensional PLLR at every 40 milliseconds to align with the gold standard emotion ratings. Delays were compensated by shifting the features forward in time. Training data were scaled into the range [0, 1] and scaling coefficients were used to normalize test data, as per [21]. The iteration number of RVM classifier and regressor was optimized on the development partition with maximum 150 iterations. In RVM-SR, the distribution of ratings on training data was evenly divided into 20 partitions for arousal and valence, selected empirically. We adopted the same post-processing including smoothing, centering and scaling as in [5]. The evaluation measure for emotion prediction is concordance correlation coefficient (CCC) [4], [5], which combines correlation coefficient and squared mean error.

4.3. Performances of PLLR features

This section compares the 25-dimensional EGEMAPS low-level descriptors, extracted using the *openSMILE* toolkit [31], with short-term PLLR features for emotion prediction using RVM. The delays were optimized, 2.4 seconds for arousal and valence for both feature sets. Post-processing was not included in this initial experiment. In Table 1, it is shown that directly applying phonetic features outperforms the commonly used acoustic features for predicting both arousal and valence.

Table 1: *Performance in CCC for short-term EGEMAPS and PLLR*

	Arousal	Valence
EGEMAPS LLDs	0.384	0.122
Short-term PLLR	0.441	0.124

However, the short-term PLLR might suffer from frame-to-frame variability as the short-term acoustic features do. To examine the effect of this variability, we calculated 5 functionals of the short-term PLLR features, i.e. mean, standard deviation, 20% percentile, 80% percentile and the range of 20%-80% percentiles, leading to 295-dimensional features with different window sizes at 40ms basis. In addition, results from a smoothed PLLR feature set via mean filter with different window sizes are presented in Fig. 2.

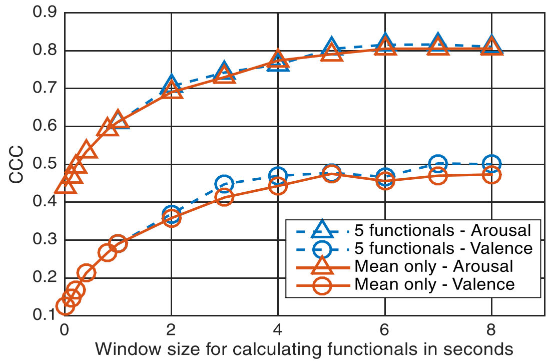
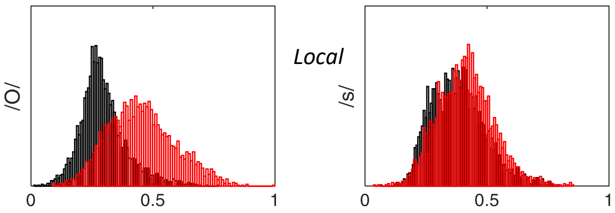


Fig. 2. *Smoothed PLLR vs 5 functionals calculated from PLLR*

There are two interesting observations from Fig. 2. The first is the smoothed PLLR features performed equally well when comparing with 5 functionals. This suggests that the improvement in performances from functionals mainly derives from the smoothed mean of the PLLR features. The second observation is the boost in performance when considering larger window sizes for both arousal and valence. This suggests that frame-to-frame variability in the short-term PLLR features can be mitigated in a similar manner to other functional features.

To look more closely into the differences between the short-term PLLR (referred to as "local") and the smoothed PLLR using 7s (referred to as "global"), a RVM classifier $C_{l=1}^{h=N}$ was trained to identify the most relevant phoneme, which was /*O*/ and /*s*/ for arousal and valence respectively. The distributions for two classes are shown in Fig. 3.



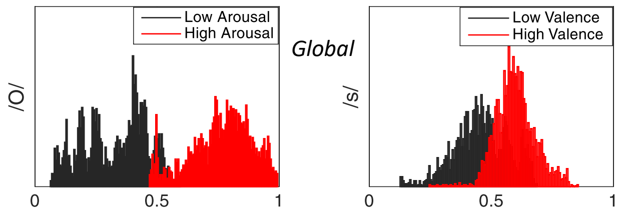


Fig. 3. *Distributions (around 7000 frames) of short-term and long-term PLLR for /O/ of high and low arousal and for /s/ of high and low valence, across all 9 speakers on training data.*

It can be clearly seen that the 'global' (long-term) PLLR features are more emotionally discriminative than the 'local' (short-term) PLLR features for both arousal and valence. The good discrimination of /*O*/ is consistent with previous literature showing that vowels are effective for emotional speech [11], [12], [13].

4.4. RVM SR vs RVM

Given the effectiveness and discriminative effect of the PLLR features, we compared the 88-dimensional EGEMAPS features and the 59-dimensional smoothed PLLR features using Support Vector Regression (SVR), RVM and RVM-SR, alongside with fusion of the two feature sets. Post-processing was included in all systems. The PLLR features were smoothed using a mean filter of 7s. The delays were optimized for the EGEMAPS (2.8 secs for arousal and 2.4 secs for valence) and PLLR features (2.8 secs for arousal and 4.4 secs for valence). Notice that in RVM-SR (*Staircase 2*) with 20 partitions, only 10 probabilistic outputs from 10 low-high arousal/valence classifiers were used for regression training.

Table 2: *Comparison of EGEMAPS and PLLR features using SVR, RVM and RVM-SR in CCC. Feature-level fusion using RVM was also compared with fusion at different stages using RVM-SR.*

		Arousal	Valence
SVR	EGEMAPS (Baseline [5])	0.796	0.455
	PLLR	0.838	0.438
RVM	EGEMAPS	0.794	0.430
	PLLR	0.821	0.473
	feature-level fusion	0.848	0.502
RVM-SR *(Staircase 2)*	EGEMAPS	0.794	0.286
	PLLR	0.846	***0.508***
	feature-level fusion	0.860	0.463
	classifier-level fusion	0.849	0.437
	score-level fusion	***0.861***	0.500

The SVR results were generated using scripts provided by the challenge [5]. It is shown in Table 2 that RVM with EGEMAPS features has comparable performance to the baseline in [5] for arousal while being slightly lower for valence. The phonetic PLLR features outperformed EGEMAPS features for arousal using both RVM and SVR, and for valence using RVM. This signals the promise for inclusion of phonetic features for emotion prediction.

For EGEMAPS features, applying RVM-SR performed equally to RVM for arousal but gave much lower valence results. By large contrast, RVM-SR with PLLR features produced further improvements over RVM, achieving similar performances as feature-level fusion of EGEMAPS and PLLR using RVM. This performance was achieved using only 10 probabilistic outputs from RVM classifiers compared with 147-dim concatenated features. This confirms the effectiveness of RVM-SR for exploiting the discriminative capability of phonemes, as shown in Fig. 3.

However, despite slight improvements for arousal, fusion of the two feature sets at the three different stages within RVM-SR *(Staircase 2)* did not provide any significant differences in valence results. This may not be surprising however, given that EGEMAPS gave a weak CCC of 0.286 for valence in RVM-SR (Table 2).

A comparison of the four proposed staircases was conducted in Fig. 4. Again, the EGEMAPS and PLLR features were fused at three different levels. It can be seen that the proposed additional staircases performed comparably or better than *Staircase 1*, especially for *Staircase 2*, which allows more discriminative data for training high vs low classifier pairs.

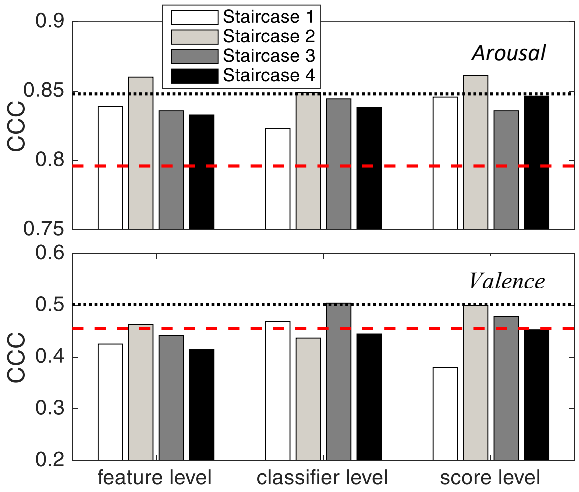


Fig. 4. *Comparison of different types of staircases. The red dashed line denotes the baseline performances in [5], while the black dashed line represents feature-level fusion of EGEMAPS and PLLR using RVM without staircase regression.*

5. CONCLUSIONS

Direct application of phonetic features has not been explored to date in speech emotion recognition; however by introducing Phone Log-Likelihood Ratio (PLLR) features to predict arousal and valence on a frame-by-frame basis, they show great promise, opening new research possibilities in this area.

Short-term and long-term PLLR features were evaluated, and the latter showed significant improvements due to the mitigation of frame-to-frame variability. In the proposed RVM-SR framework, PLLR features achieved the best valence performance of 0.508, and best arousal performance of 0.861 when fusing EGEMAPS features, yielding an 8.2% and 11.6% relative improvement in CCC over single-feature systems [5] for arousal and valence respectively. This suggests that the discriminative power of phone LLR features are well exploited in proposed RVM-SR framework. Moreover, this confirms our speculation that PLLR features are complementary to widely-used acoustic features. Future work involves further exploiting the discriminative power of phonemes by constructing staircase classifiers for each phoneme for continuous emotion prediction.

6. ACKNOWLEDGEMENT

This work is partly funded by Data61, CSIRO. The authors would like to thank their colleagues Mr. Sarith Fernando and Mr. Saad Irtza for their help with the PLLR features and helpful discussions.

7. REFERENCES

[1] H. Gunes and B. Schuller, “Categorical and dimensional affect analysis in continuous input: Current trends and future directions,” *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[2] R. Cowie, G. McKeown, and E. Douglas-Cowie, “Tracing Emotion,” *International Journal of Synthetic Emotions*, vol. 3, no. 1, pp. 1–17, 2012.

[3] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, “AVEC 2012: the continuous audio/visual emotion challenge,” *Proc. 14th Int’l Conf. Multimodal Interaction Workshops*, pp. 449–456, 2012.

[4] F. Ringeval, B. Schuller, S. Jaiswal, M. Valstar, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, “AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data,” in *Proceedings of the 5th International Workshop on AVEC, ACM MM*, 2015, pp. 3–8.

[5] M. Valstar, J. Gratch, F. Ringeval, M. T. Torres, S. Scherer, and R. Cowie, “AVEC 2016 – Depression , Mood , and Emotion Recognition Workshop and Challenge,” in *Proceedings of the 6th International Workshop on AVEC, ACM MM*, 2016.

[6] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. A. E, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[7] V. Sethu, J. Epps, and E. Ambikairajah, “Speech based emotion recognition,” in *Speech and Audio Processing for Coding, Enhancement and Recognition*, 2015, pp. 197–228.

[8] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, “On the influence of phonetic content variation for acoustic emotion recognition,” *Lecture Notes in Computer Science*, vol. 5078 LNCS, pp. 217–220, 2008.

[9] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 emotion challenge.,” *INTERSPEECH*, pp. 312–315, 2009.

[10] S. Mariooryad and C. Busso, “Compensating for speaker or lexical variabilities in speech for emotion recognition,” *Speech Communication*, vol. 57, pp. 1–12, 2014.

[11] V. Sethu, E. Ambikairajah, and J. Epps, “Phonetic and speaker variations in automatic emotion classification,” in *INTERSPEECH*, 2008, pp. 617–620.

[12] C. Lee, S. Yildirim, and M. Bulut, “Emotion recognition based on phoneme classes.,” in *INTERSPEECH*, 2004, pp. 889–892.

[13] F. Ringeval and M. Chetouani, “A vowel based approach for acted emotion recognition,” *INTERSPEECH*, pp. 2763–2766, 2008.

[14] D. Bitouk, R. Verma, and A. Nenkova, “Class-level spectral features for emotion recognition,” *Speech Communication*, vol. 52, pp. 613–625, 2010.

[15] A. Origlia, F. Cutugno, and V. Galatà, “Continuous emotion recognition with phonetic syllables,” *Speech Communication*, vol. 57, pp. 155–169, 2014.

[16] S. O. Caballero-Morales, “Recognition of emotions in Mexican Spanish speech: An approach based on acoustic modelling of emotion-specific vowels,” *The Scientific World Journal*, vol. 2013, pp. 13–16, 2013.

[17] B. Vlasenko, D. Prylipko, R. Böck, and A. Wendemuth, “Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications,” *Computer Speech and Language*, vol. 28, no. 2, pp. 483–500, 2014.

[18] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, “Vocal biomarkers of depression based on motor incoordination,” in *Proceedings of the 4th ACM International Workshop on AVEC, ACM MM*, 2013, pp. 41–47.

[19] J. Williamson, T. Quatieri, and B. Helfer, “Vocal and facial biomarkers of depression based on motor incoordination and timing,” in *Proceedings of the 4th International Workshop on AVEC, ACM MM*, 2014.

[20] N. Cummins, “Automatic Assessment of Depression from Speech: Paralinguistic Analysis, Modelling and Machine Learning,” *PhD Thesis, UNSW Australia*, 2016.

[21] Z. Huang, B. Stasak, T. Dang, K. Wataraka Gamage, L. Phu, V. Sethu, and J. Epps, “Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016,” in *Proceedings of the 6th International Workshop on AVEC, ACM MM*, 2016.

[22] I. Mporas, T. Ganchev, and N. Fakotakis, “Phonetic Segmentation of Emotional Speech With Hmm-Based Methods,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 24, no. 7, pp. 1159–1179, 2010.

[23] P. Roach, “Techniques for the phonetic description of emotional speech,” *ISCA Workshop on Speech and Emotion*, pp. 53–59, 2000.

[24] M. Diez, A. Varona, and M. Penagarikano, “On the use of phone log-likelihood ratios as features in spoken language recognition,” *(SLT), 2012 IEEE*, 2012.

[25] Z. Huang, T. Dang, N. Cummins, B. Stasak, L. Phu, V. Sethu, and J. Epps, “An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction,” in *Proceedings of the 5th International Workshop on AVEC, ACM MM*, 2015.

[26] M. Sánchez, “Frame-Level Features Conveying Phonetic Information for Language and Speaker Recognition,” *PhD Thesis, University of The Basque Country*, 2015.

[27] M. Schmitt, F. Ringeval, and B. Schuller, “At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech,” in *INTERSPEECH*, 2016.

[28] M. Tipping, “Sparse Bayesian learning and the relevance vector machine,” *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[29] F. Ringeval, A. Sonderegger, J. Sauser, and L. D, “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions,” in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).*, 2013, pp. 1–8.

[30] P. Schwarz, “Phoneme recognition based on long temporal context,” *PhD Thesis, Brno University of Technology*, 2008.

[31] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.