



# Role of Regularization in the Prediction of Valence from Speech

Kusha Sridhar, Srinivas Parthasarthy, Carlos Busso

Multimodal Signal Processing(MSP) lab, Department of Electrical and Computer Engineering  
The University of Texas at Dallas, Richardson TX 75080, USA

Kusha.Sridhar@utdallas.edu, sxp120931@utdallas.edu, busso@utdallas.edu

## Abstract

Regularization plays a key role in improving the prediction of emotions using attributes such as arousal, valence and dominance. Regularization is particularly important with *deep neural networks* (DNNs), which have millions of parameters. While previous studies have reported competitive performance for arousal and dominance, the prediction results for valence using acoustic features are significantly lower. We hypothesize that higher regularization can lead to better results for valence. This study focuses on exploring the role of dropout as a form of regularization for valence, suggesting the need for higher regularization. We analyze the performance of regression models for valence, arousal and dominance as a function of the dropout probability. We observe that the optimum dropout rates are consistent for arousal and dominance. However, the optimum dropout rate for valence is higher. To understand the need for higher regularization for valence, we perform an empirical analysis to explore the nature of emotional cues conveyed in speech. We compare regression models with speaker-dependent and speaker-independent partitions for training and testing. The experimental evaluation suggests stronger speaker dependent traits for valence. We conclude that higher regularization is needed for valence to force the network to learn global patterns that generalize across speakers.

**Index Terms:** regularization, dropout, detection of valence.

## 1. Introduction

Automatic affect recognition plays a key role in *human computer interactions* (HCIs) and behavioral problems, which are multimodal, subtle and complex. Speech conveys affective information through linguistic and paralinguistic cues, which are extremely important in understanding the meaning of a sentence. Recognizing emotional traits in speech is still a challenging task. Psychology suggests two major theories to describe emotions - (1) basic emotions (i.e., categorical) [1, 2] and (2) core emotions (i.e., attributes) [3]. While previous studies have mostly focused on recognizing basic emotions, detecting emotional attributes has many advantages. During everyday interactions, people exhibit rather complex affective behaviors that cannot be easily described with basic emotions. According to the core affect theory [4], the majority of expressive behaviors can be described with few attributes, which are not independent, but related in a systematic manner. The most common attributes are arousal (calm versus active), valence (unpleasant versus pleasant) and dominance (weak versus strong).

Previous studies have reported competitive performance for detecting arousal and dominance from speech. However, the performance for valence is commonly lower, as acoustic features are less discriminative for this task [5]. This is an important problem since valence plays a key role in many behav-

ioral problems (e.g., detecting trauma, depression, shock). Several studies have tried to improve the prediction of emotional attributes, including valence, using different approaches. Studies have presented systematic analyses to explore discriminative acoustic features for valence [5], considered features from other modalities [6], or modeled contextual information [7]. Other studies have relied on *multitask learning* (MTL) to regularize a DNN by jointly detecting multiple attributes [8]. We hypothesize that acoustic cues for valence are more speaker-dependent than acoustic cues for arousal and dominance. As a result, optimal models for predicting valence require higher regularization than models for dominance and arousal to learn consistent trends that generalize across speakers. This analysis explores this hypothesis with controlled evaluations, focusing our effort on valence.

The first evaluation considers the role of the dropout rate on the performance of the DNN. Dropout is an effective approach to regularize DNNs, where nodes in the layers are randomly turned off during training [9]. We train regression models for emotional attributes by systematically increasing the dropout probability. We record the network performance in terms of *concordance correlation coefficient* (CCC) for valence, arousal, and dominance under different experimental conditions. The analysis suggests that the optimal dropout rate for valence is consistently higher than the optimal dropout rates for arousal and dominance. To understand the underlying reasons for the need of higher regularization for valence, the second evaluation compares the difference in performance of emotional regressors trained with speaker-dependent and speaker-independent train and test partitions. In the speaker-independent condition, we avoid overlap of data from the same speaker in the train and test partitions. In the speaker-dependent conditions, we add half of the data from every test speaker into the train set. While the relative improvements in CCC for arousal and dominance were less than 4%, the CCC values for valence have relative improvements over 28%, when the train set includes data from the target speakers. The results indicate that acoustic cues for valence are more speaker-dependent than acoustic cues for arousal or dominance. Higher regularization allows the DNN to identify trends that generalize across speakers, improving the performance for valence. This analysis has important implications in the training of machine-learning solutions, suggesting that special considerations should be forethought when training valence models.

## 2. Related Work

### 2.1. Detecting Valence from Speech

Valence indicates the pleasure level, describing whether the behavior is emotionally positive or negative. It plays an important role in many areas including health-care (e.g., mood disorders), customer service (call center), and education (e.g., frustration level). Although studies on speech emotion recognition have made important advances, the prediction of valence from

This work was funded by NSF CAREER award IIS-1453781.

speech is still a challenging problem. The performance for valence is commonly worse than the one for other emotional attributes [10–12]. Wu et al. [13] combined spectral and prosodic features to improve the emotion regression models. The correlation between the predictions and ground-truth labels was significantly lower for valence than for arousal and dominance. Similar results were observed by Parthasarathy and Busso [8]. Neumann and Vu [14] proposed regression models for arousal and valence for mono-lingual and multi-lingual evaluations. The performance was lower for valence. Furthermore, the performance for valence decreased for multi-lingual evaluation. This trend was not observed for arousal, where the results improved.

Busso and Rahman [5] studied acoustic properties that describe valence. They trained separate regression models using different groups of acoustic features, showing that spectral features and features from the fundamental frequency are the most discriminative cues for valence. Liscombe et al. [15] showed that acoustic features such as spectral tilt, type of phrase accent, and boundary tone in speech are useful to discriminate valence. Goudbeek et al. [16] studied the influence of emotional attributes in the position of the formants. The study found that the mean of the second formant was higher for sentences with positive valence than for sentences with negative valence.

## 2.2. Psychological Perspective on the Expression of Valence

Various studies in psychology have shown the contribution of valence in the expression of emotion. Barrett [17] hypothesized that valence is one of the basic components in most expressive behaviors. This study further showed that people differ in the externalization of pleasure or displeasure, arguing that the variations in the expression of valence depends on the individual’s appraisal of the situations. Feldman [18] showed that people generally weigh valence more than arousal when making judgments about their mood. By analyzing self-reported mood, Feldman [19] also reported that the variance for valence was almost twice the variance for arousal. These studies indicate that the expression of valence has more idiosyncratic characteristics than other emotional attributes.

## 2.3. Regularization in DNNs

DNNs have significantly improved the performance of speech emotion recognition systems [8, 20–22]. As the number of parameters increases for more sophisticated networks, it is important to avoid overfitting. Several regularization methods have been proposed to improve the generalization of the models. Regularization aims to learn trends that generalize across conditions [23], penalizing patterns in the training set that are too specific. Although new emotional databases are made available [24], their sizes are still small compared to resources in other speech fields. Therefore, regularization is very important for this task. Commonly used regularization techniques are data augmentation, early stopping, dropout layers and weighted penalties such as L1 and L2.

This study uses dropout as a form of regularization [9]. Dropout regularizes the network by randomly turning off  $p$  percent of the nodes in the network for each epoch. Dropout is similar to training a smaller network on every iteration. It reduces co-adaptation and inter-dependencies among nodes, which leads to more general models. Srivastava et al. [9] showed that dropout increases the performance of the network under supervised learning tasks for speech recognition. This study explores optimal dropout rates for detecting valence, comparing them with the optimal rates for arousal and dominance.

# 3. Resources

## 3.1. The MSP-Podcast Database

The study uses the MSP-Podcast corpus [24], which is a collection of emotionally rich spontaneous speech samples from various podcasts collected from audio-sharing websites. These audio podcasts contain conversations about various topics including politics, business, science, technology, lifestyle, sport, movie and economy. The podcasts are segmented into speaking turns containing no background music or overlapped speech, following the procedure described in Lotfian and Busso [24]. The study uses version 1.0 of the MSP-Podcast corpus, which consists of 20,045 emotionally annotated speech segments (30h43m). The test data has 6,069 samples from 50 speakers and the validation data has 2,226 samples from 15 speakers. The rest of the corpus is included in the training set. The data partition attempts to create datasets with minimal speaker overlap between sets. The corpus is annotated by at least five evaluators, using a similar crowdsourcing protocol introduced by Burmania et al. [25]. This study uses the annotations for valence, arousal and dominance. The raters completed the evaluation using a seven Likert-type scale for each attribute. The ground truth for a speaking turn is the average across the scores provided by the annotators.

## 3.2. Acoustic Features

This study uses the feature set proposed for the computational paralinguistics challenge in Interspeech 2013 [26]. The features are extracted with OpenSmile [27]. First, we extract *low level descriptors* (LLDs) such as energy, spectral and cepstral features. From the LLDs, we estimate statistical measures such as mean, moments and regression coefficients, creating a vector with 6,373 features for each speech segment.

# 4. Experimental Framework and Results

The analysis explores the role of the dropout rate in detecting emotional attributes (Sec. 4.1), showing the need for higher regularization for valence. Section 4.2 demonstrates that this finding is consistently observed across different DNN structures. Section 4.3 explores the underlying reasons behind this finding.

The prediction of emotional attributes is formulated as a regression problem implemented with DNNs. The analysis explores DNNs with different layers and different nodes per layer. We use *rectified linear unit* (ReLU) at the hidden layers and linear activations for the output layer. The DNNs are trained with *stochastic gradient descent* (SGD) with momentum of 0.9, and a learning rate of 0.001. The models are trained with CCC as the cost function, since recent studies have shown that CCC gives better performance for detecting emotional attributes [10]. CCC captures not only the correlation between the true emotional labels and their estimates, but also the difference in their means. CCC is also the performance metric in our analysis. The DNNs are trained with batch normalization for the hidden layers, as it reduces covariance shift. Batch normalization helps normalize the output of each layer, leading to faster training and better gradient flow. Furthermore, studies have shown that batch normalization leads to better performance in deep networks [22, 28, 29]. We maximize the CCC on the validation set for each setting (number of nodes per hidden layer, number of hidden layers, different dropout probabilities, target emotional attribute). We train the models for 1,000 epochs for the analysis.

The input to the DNNs is the 6,373D acoustic feature vector (Sec. 3.2). The features are standardized using the mean and standard deviation values estimated over the training samples.

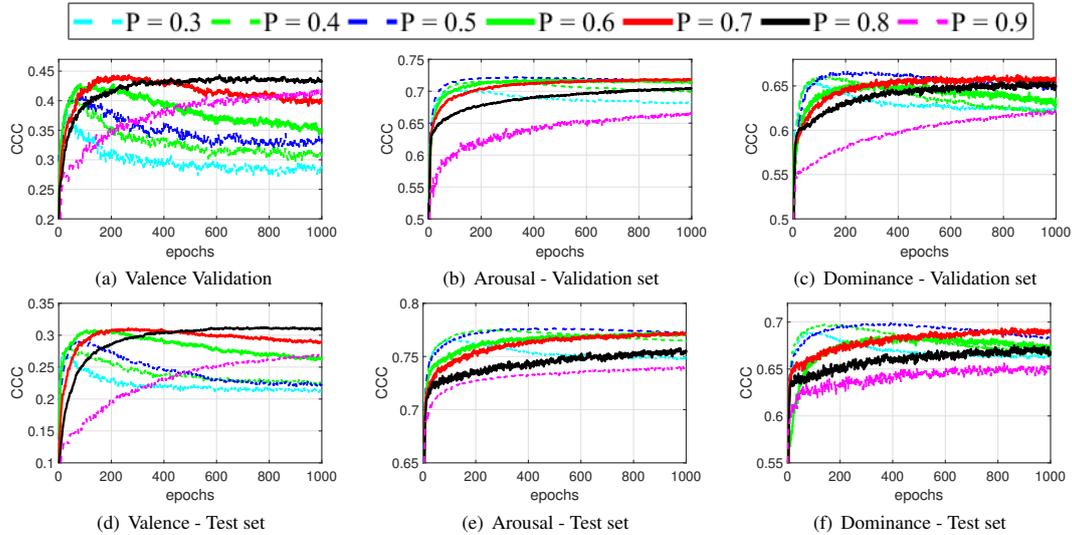


Figure 1: CCC scores on the validation and test sets over 1,000 epochs for different dropout rates.

We attenuate all the features for which their values deviate more than three standard deviations from the mean. The output of the DNNs is the predicted score for valence, arousal or dominance.

#### 4.1. Performance as a Function of Dropout Rate

The first evaluation considers the role of the dropout rate on the performance of DNNs. A higher rate implies higher regularization. This analysis considers a DNN with two layers, each with 256 nodes. We train separate emotion regression models by changing the dropout rate of the hidden layers from  $p = 0.0$  to  $p = 0.9$  in increments of 0.1 (i.e., 10 models per emotional attribute). We report the changes in CCC values in the validation and test sets as a function of the number of epochs.

Figure 1 reports the changes in predicting CCC for each value of  $p$  as the model progresses through the epochs. Figures 1(a)-1(c) report the performance for the validation set, and Figures 1(d)-1(f) report the performance for the test set. The plots show a clear difference for the optimal value of  $p$  needed for valence, arousal, and dominance. The CCC curves for valence reach a peak performance for dropout rates  $p \in \{0.7, 0.8\}$ , whereas the peak performance for arousal and dominance is achieved with relatively lower dropout rates ( $p \in \{0.4, 0.5\}$ ). In the test set, the worse performance for arousal and dominance is with either  $p = 0.8$  or  $p = 0.9$ . In contrast, the performance for valence with  $p = 0.8$  is very competitive in both sets (Figs. 1(a) and 1(d)).

#### 4.2. Performance with Other Network Configurations

The previous section demonstrates the need for a higher dropout rate for valence. This section shows that this finding also holds for other DNN configurations. We consider DNNs with more layers and/or with more nodes per layers, evaluating the optimal value for the dropout rate for valence, arousal and dominance.

First, we train the models with two layers but with different numbers of nodes per layer (256, 512, and 1,024). Table 1 compares the average CCC predictions for  $p = 0.5$  and  $p = 0.7$  over 10 trials with different random initialization of the parameters. For the validation set, we report the best performance obtained across the 1,000 epochs. The optimal epoch number is used to evaluate the results on the test set. Notice that the early

Table 1: CCC values achieved on the validation and test sets for dropout rates equal to  $p = 0.5$  and  $p = 0.7$ . The DNN has two layers. <sup>†</sup> indicates that one dropout rate leads to significantly better results than the other dropout rate.

Attributes	Nodes	Validation		Test	
		$p = 0.5$	$p = 0.7$	$p = 0.5$	$p = 0.7$
Valence	256	0.4258	0.4485	0.2903	0.3102 <sup>†</sup>
	512	0.4220	0.4383	0.2870	0.3080 <sup>†</sup>
	1,024	0.4159	0.4323	0.2841	0.3009 <sup>†</sup>
Arousal	256	0.7167	0.7080	0.7733 <sup>†</sup>	0.7577
	512	0.7151	0.6980	0.7717 <sup>†</sup>	0.7525
	1,024	0.7135	0.6926	0.7691 <sup>†</sup>	0.7472
Dominance	256	0.6624	0.6493	0.6936 <sup>†</sup>	0.6733
	512	0.6597	0.6337	0.6902 <sup>†</sup>	0.6617
	1,024	0.6557	0.6241	0.6888 <sup>†</sup>	0.6523

stopping criterion may be different across trials. Regardless of the number of layers in the network, the results on Table 1 agree with the patterns observed in Figure 1. For valence, the performance is higher for higher dropout rate ( $p = 0.7$ ). A one-tailed t-test over the 10 trials indicates that the differences are statistically significant ( $p\text{-value} \leq 0.001$ ). For arousal and dominance, using  $p = 0.5$  leads to better performance than with  $p = 0.7$ . The results are statistically significant (one-tailed t-test over 10 trials, asserting significance if  $p\text{-value} \leq 0.001$ ).

We also evaluate the optimal dropout rate for DNNs implemented with different numbers of layers (2, 4, and 6 layers). For each of these cases, the number of nodes per layer is 256. The value for the optimal value of  $p$  is obtained on the validation set. Figure 2 presents the results as a function of the number of layers. A key result in Figure 2 is that, for a fixed number of layers, the optimal value for  $p$  is always higher for valence than for arousal and dominance. It is interesting that the DNNs for arousal and dominance are optimized with the same value for  $p$ . The figure shows that the value for the optimal  $p$  consistently decreases for all the attributes as we increase the depth of the DNN. However, valence still requires higher regularization.

Figure 3 reports the optimal dropout rate as we increase the number of nodes per layer (256, 512, and 1024). The value for the optimal  $p$  is also obtained on the validation set. We present results for DNNs with two and six layers. The figures follow

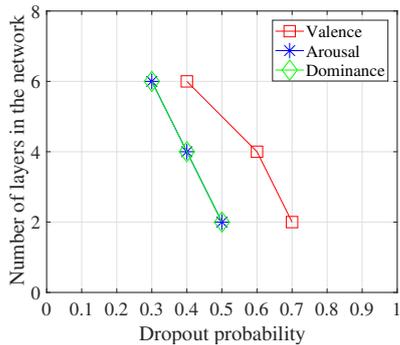


Figure 2: Optimal dropout rate as a function of the depth of the network. Each layer has 256 nodes.

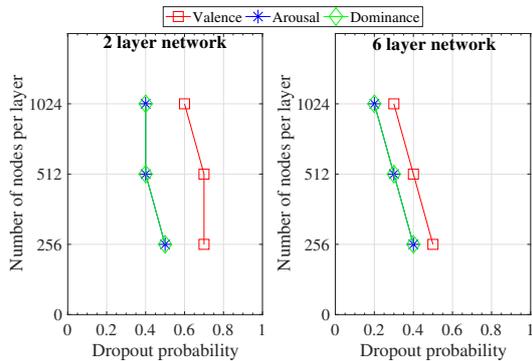


Figure 3: Optimal dropout rate as a function of the number of nodes per layer. The DNN is trained with either 2 or 6 layers.

similar trends to the ones observed in Figure 2: (1) valence requires higher regularization than arousal and dominance, and (2) the optimal  $p$  tends to decrease as the DNN is implemented with more nodes. The analysis confirms the need for higher regularization for valence, regardless of the structure of the DNN.

#### 4.3. Reasons Behind Higher Regularization for Valence

We hypothesize that DNNs for valence require higher regularization because of the speaker-dependent nature of acoustic cues for this attribute. Higher regularization helps the network to learn more discriminative features that are consistently observed across all speakers, reducing the weights of speaker-dependent trends that are too specific. To explore this hypothesis, we compare DNNs trained with speaker-dependent and speaker-independent train and test partitions. The key idea in this evaluation is to quantify the benefits of training emotional models with data from speakers in the test set. If the gap in performance between these conditions is big, we conclude that emotional traits used by the DNNs are more speaker-dependent.

The test set is split in two groups, where half of the samples from every speaker is placed in each group. One group is used as the new test set. For the speaker-independent condition, the second group is discarded. For the speaker-dependent condition, however, the second group is added to the training set. Therefore, the training set has speech samples for each speaker in the training set. Note that this evaluation is only for analysis purposes, as emotion recognition systems should be trained with speaker independent partitions.

We conduct this evaluation with DNNs with two hidden lay-

Table 2: Comparison of CCC values between speaker independent and dependent conditions. The DNN is trained with two layers. The column ‘Gain’ shows the relative increase by training with data from the target speakers.

Attributes	Nodes	Speaker Independent	Speaker Dependent	Gain (%)
		Test	Test	Test
Valence	256	0.2906	0.3761	29.42
	512	0.2835	0.3686	30.01
	1,024	0.2880	0.3600	28.57
Arousal	256	0.7712	0.7885	2.24
	512	0.7720	0.7813	1.20
	1,024	0.7688	0.7800	1.45
Dominance	256	0.6901	0.7051	2.17
	512	0.6837	0.7052	3.14
	1,024	0.6782	0.7005	3.28

ers and with 256, 512 or 1,024 nodes per layers. The DNNs are trained with dropout rate  $p = 0.5$  for arousal and dominance, and  $p = 0.7$  for valence, since these values provide better performance for the speaker-independent condition on the validation set. Table 2 shows the CCC values obtained with speaker-dependent and speaker-independent training partitions. The last column shows the relative gain observed with the speaker-dependent condition. Valence obtains relative improvements over 28%. For arousal and dominance, the relative improvement is below 4%. With speaker-dependent training, the DNNs learn acoustic cues that are characteristic of the target speakers, leading to important performance gains for valence. These results are not observed with arousal and dominance. These results indicate that stronger regularization is required to learn more general acoustic features that are consistent across speakers.

## 5. Conclusions

This paper analyzed the optimal value of the dropout rate for emotion detection systems for valence, arousal and dominance. The systematic analysis demonstrated the need for higher regularization for valence, as its optimal dropout rate was found to be higher than the ones for arousal and dominance ( $p \in \{0.7, 0.8\}$  for valence,  $p \in \{0.4, 0.5\}$  for arousal and dominance). This finding was consistently observed for different configurations of our DNNs (different number of layers, different number of nodes per layers). The results suggest that heavy regularization is needed between the layers for detecting valence. We hypothesized that discriminative acoustic features for detecting valence vary across speakers. We evaluated this hypothesis by training regression models with speaker-independent and speaker-dependent training partitions. Valence was the only emotional attribute that benefited from training the models with data from speakers included in the test set, with relative gains above 28%. Higher regularization forces the models to identify more general acoustic patterns that are observed across speakers.

The analysis in this study has important implications for emotion recognition, where regression models for valence, arousal and dominance are commonly implemented with the same network configurations, including their dropout rates. Our analysis suggests that treating all the emotional attributes equally is a mistake, as they have important differences. For our future work, we want to extend the analysis with other regularization techniques. We will evaluate the results on other emotional databases. Finally, we will also explore whether our finding is also observed in more complex deep learning frameworks such as generative adversarial networks (GANs), recurrent neural networks (RNNs) or ladder networks.

## 6. References

- [1] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
- [2] G. Colombetti, "From affect programs to dynamical discrete emotions," *Philosophical Psychology*, vol. 22, no. 4, pp. 407–425, August 2009.
- [3] H. Schlosberg, "Three dimensions of emotion," *Psychological review*, vol. 61, no. 2, p. 81, March 1954.
- [4] J. Russell, "Core affect and the psychological construction of emotion," *Psychological review*, vol. 110, no. 1, pp. 145–172, January 2003.
- [5] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.
- [6] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, April-June 2011.
- [7] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1983–1986.
- [8] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.
- [9] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, June 2014.
- [10] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.
- [11] S. Parthasarathy, R. Cowie, and C. Busso, "Using agreement on direction of change to build rank-based emotion classifiers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2108–2121, November 2016.
- [12] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5205–5209.
- [13] S. Wu, T. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, May-June 2011.
- [14] M. Neumann and N. Vu, "Cross-lingual and multilingual speech emotion recognition on English and French," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5769–5773.
- [15] J. Liscombe, J. Venditti, and J. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," in *8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Geneva, Switzerland, September 2003, pp. 725–728.
- [16] M. Goudbeek, J. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1575–1578.
- [17] L. Barrett, "Valence is a basic building block of emotional life," *Journal of Research in Personality*, vol. 40, no. 1, pp. 35–55, February 2006.
- [18] L. Feldman, "Variations in the circumplex structure of mood," *Personality and Social Psychology Bulletin*, vol. 21, no. 8, pp. 806–817, August 1995.
- [19] —, "Valence focus and arousal focus: Individual differences in the structure of affective experience," *Journal of personality and social psychology*, vol. 69, no. 1, pp. 153–166, December 1995.
- [20] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli, "Hybrid deep neural network-hidden markov model (DNN-HMM) based speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 312–317.
- [21] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, Singapore, September 2014, pp. 223–227.
- [22] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5084–5088.
- [23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: The MIT Press, November 2016.
- [24] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. To appear, 2018.
- [25] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.
- [26] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Wengler, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.
- [27] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.
- [28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (PMLR 2015)*, vol. 37, Lille, France, July 2015, pp. 448–456.
- [29] H. Fayek, M. Lech, and L. Cavedon, "On the correlation and transferability of features between automatic speech recognition and speech emotion recognition," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3618–3622.