A Robust Unsupervised Arousal Rating Framework using Prosody with Cross-Corpora Evaluation

Daniel Bone, Chi-Chun Lee, Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA
{dbone@, chiclee@, shri@sipi}.usc.edu

Abstract

This paper presents an unsupervised method for producing a bounded rating of affective arousal from speech. One of the major challenges in such behavioral signal classification is the design of methods that generalize well across domains and datasets. We propose a framework that provides robustness across databases by: selecting coherent features based on empirical and theoretical evidence, fusing activation confidences from multiple features, and effectively weighting the soft-labels without knowing the true labels. Spearman's rank-correlation (and binary classification accuracy) on four arousal databases are: 0.62 (73%), 0.77 (86%), 0.70 (82%), and 0.65 (73%).

Index Terms: arousal rating, activation, unsupervised, knowledge-based, inter-rater reliability, cross-corpora

1. Introduction

Emotion guides human behavior in conscious and unconscious ways and continues to be a phenomenon that evokes multi-disciplinary scholarly interest. Efficient human communication is dependent on multi-modal reception, processing, and transmission of affective cues. Imagine, for example, the benefits of adapting one's behavior based on a partner's sometimes subtle emotional signals. The study of emotion ranges from the fields of psychology and sociology, biology, engineering, linguistics, and even consumer research and risk analysis.

Affect is a complex process that is considered essential to developing psychological theory and interpreting experimental results, and speech is a critical medium for studying affect. Pollermann [1] has argued that prosody is essential to fully understand cognition and emotion. In many cases, prosodic correlates of affect are used as variates for analyzing human behavior. For example, pitch has been used as a measure of arousal to analyze depressed patient intervention outcomes [2], to produce a visual depiction of arousal for negotiators [3], and to analyze coping power in patients with breast cancer [1].

While pitch is an essential tool for conveying paralinguistic information, a speaker may display emotions through other mechanisms and modalities. Juslin and Scherer (2005) suggested that interactions among higher-order variables may reflect combinations of measures more closely related to human perception than low-order variables. For example, 'vocal effort' may be a combination of acoustic features such as voice intensity and high-frequency energy [4]. Integrating aggregate-variates into the experimental design would provide increased modeling capability and potential robustness, but such applications with natural affect often do not contain associated labeled data, limiting the relevance of supervised algorithms. Additionally, speaker-normalization, while a fundamental signal processing tool for increasing generalizability, is not observed to be universally applied when it is preferable or even necessary. Thus, human behavior researchers are in need of robust measures of affective phenomena that are easily transferrable across corpora.

Engineers have focused on the importance of emotion for system design. Affective computing work on emotional speech has been motivated by human-machine interaction [5], speech recognition, and speaker identification [6]. As a result, the emphasis has been on optimizing classification results for a given domain (database).

A new emphasis in emotion recognition is emerging in which the goal is cross-corpus robustness such that engineering methodologies can be more readily applied to unlabeled data. Arousal recognition with speech has found success but nevertheless with corpora dependent variability– the success is generally much lower for identifying valence from speech. Eyben et al. [7] and Schuller [8] have demonstrated above-chance accuracies in dimensional emotion classification without any normalization across corpora, but Eyben et al. have noted that the methodology still needed to mature. In order to achieve higher performance, it is necessary to consider speaker normalization to adapt to a new corpus and person. Schuller et al. demonstrated that a 1,406-dimensional feature vector trained with speaker normalization achieves much higher accuracies across corpora [9]. Affective encoding in speech is a noisy process due to contextual factors and speaker and corpus variability; consequently there has been little agreement across corpora on optimal feature sets and parameters.

In this work, we seek to connect the goals and constraints of psychological and engineering research by providing a robust automatic, unsupervised knowledge-

Table 1: *Description of emotional corpora and arousal labels.*

Corpus	**Style**	**Emotion**	**Label**	$-$	$+$	Neu	**Total**	**Speakers**	**Setting**	**Language**
IEMOCAP	spontaneous & scripted	acted	ordinal	2579	4304	(1112)	6883	10 (5f,5m)	studio	English
EMA	read	acted	categorical	408	338	221	967	3 (1f,2m)	studio	English
emoDB	read	acted	categorical	189	267	79	535	10 (5f,5m)	studio	German
VAM	spontaneous	natural	continuous	502	445	N/A	947	47 (32f,15m)	noisy	German

based framework for producing ratings of arousal while incorporating multiple prosodic features. Guided by the hundreds of empirical studies summarized by Juslin and Scherer [4], we select an interpretable feature set that is both coherent across corpora and robust in automatic extraction, requiring limited assumptions on the data. The chosen features are median log-pitch, median intensity, and HF500 (similar to spectral slope). Next, we obtain an arousal soft-rating from each feature in reference to a speaker's neutral state features– the algorithm requires a small amount of neutral speech per speaker. Lastly, we treat the ratings as coming from noisy labelers and fuse the labels weighted by their correlation with the average rating as in [10]. This provides robustness when one of the three features is corrupted, but the others are unaltered. We also demonstrate that our methodology is effective across two languages, German and English.

The rest of this paper is organized as follows: the databases and methodology are presented in Section 2 and Section 3, the experimental results are discussed in Section 4, and the conclusion is made in Section 5.

2. Databases

We have experimented with four publicly available databases comprising scripted and spontaneous emotional speech in German and English as well as natural German emotional speech. The databases are IEMOCAP, EMA, emoDB, and VAM (details in Table 1).

2.1. Acted Emotional Speech Copora

IEMOCAP is a database of mixed-gender dyadic interaction between actors [11]. It contains data from 5 dyads interacting in both spontaneous improvisation of hypothetical emotional scenarios (2388 utterances) and portrayal of scripted emotional content (4517 utterances).

IEMOCAP data is tagged for categorical and dimensional emotional labels by at least two raters. The target labels for the following experiments are the average activation ratings on an integer scale from 1 to 5. The data was tagged sequentially using both audio and video. We also include the data where no agreement on categorical emotions is made (28%). Ambiguous displays of emotion are known to be more difficult to identify. Data with no voiced frames were discarded (<1% of utterances).

The USC-EMA corpus consists of read, emotional speech from three trained actors performing five emotions in English - neutral, hot anger, cold anger, happy, and sad. We assume hot anger and happy to be high arousal, and sadness and cold anger to be low arousal. Data is rated by at least 4 raters. Two additional types of variability are present. The speakers are asked to repeat each sentence-emotion pair in three speaking styles: normal, loud, and fast. It is expected that such variability will have a direct effect on perceived arousal. Each speaker also has sensors on the face and tongue to support articulatory study of emotions [12, 13].

emoDB is an acted database of German emotional speech with seven emotions - neutral, happy, angry, fear, sad, bored, and disgust. The latter three are considered low arousal and the preceding three are considered high arousal. The intensity in this corpus is unreliable due to varying mouth-to-mic distance [14], and a system that relies explicitly on intensity would fail.

2.2. Natural Emotional Speech Corpus

The VAM corpus [15] is a natural emotional speech database consisting of speakers in dyadic or triadic conversations on the German TV talk-show "Vera am Mittag" (Vera at noon). There are 47 distinct speakers in the audio release and a total of 947 sentences. Each sentence was transcribed by between 7 to 16 raters within a continuous-valued scale along the dimensions valence, activation, and dominance.

3. Method

We extract select features, compute Gaussians over neutral-state features, score each utterance's features versus the corresponding neutral model, and combine scores weighted by their correlations to score means.

3.1. Knowledge-Inspired Prosody Features

In order to construct an unsupervised system that generalizes across datasets, we must have features that are both theoretically and empirically consistent. Physiological theory predicts the vocal effects of arousal– for instance, fear (high arousal) will cause muscle tension in the laryngeal folds as a sympathetic nervous system response, leading to higher pitch [16]. Empirical meta-analyses of emotional speech studies corroborate such theory, showing that median and variability of pitch, median and variability of intensity, voice quality (HF500), and speaking rate all increase with arousal. Additionally, perceptual studies have demonstrated similar correlations across natural and mood-induced emotional speech [4].

From the above features, we have chosen median pitch and intensity and voice quality (HF500), all taken from voiced frames, in an effort to increase orthogonality and obtain robust feature extraction. HF500 is a voice quality, spectral-slope measure computed as the to-

Table 2: *Spearman's rank-correlation (* **Sp.** ρ*), fusion weights (***w***), and binary classification accuracy for the weighted-fusion (W-fusion) arousal rating. For VAM, there are two speaker normalization methods: neutral (NN) and global (GN).*

		F0$_{med}$		**HF500**		**INT**$_{med}$		**UW-fusion**	**W-fusion**	**Unweighted**
Corpus	**Style**	**Sp.** ρ	**w**	**Sp.** ρ	**w**	**Sp.** ρ	**w**	**Sp.** ρ	**Sp.** ρ	**Avg. Recall**
IEMOCAP	spont. & scripted	0.51	0.78	0.46	0.79	0.59	0.91	0.62	0.62	73%
EMA	read	0.62	0.74	0.73	0.83	0.74	0.91	0.76	0.77	86%
emoDB	read	0.81	0.91	0.71	0.93	-0.49	-0.02	0.62	0.70	82%
VAM NN	spontaneous	0.60	0.67	0.41	0.77	0.66	0.68	0.66	0.65	73%
VAM GN	spontaneous	0.56	0.66	0.38	0.79	0.61	0.67	0.63	0.63	70%

tal energy above 500Hz over the lower-frequency energy. Speaking rate is desired but assumes accurate ASR tools or manual transcription are in place for a given corpus. Given this configurable framework, it could easily be inserted if available. Pitch and intensity are extracted using Praat [17], and HF500 is obtained using both pitch and the original audio in Matlab. Pitch is log-transformed since it has been shown that pitch is log-normal [18].

3.2. Speaker Normalization Model

Inter-speaker variability is repeatedly seen to impede emotion classifier performance. In this unsupervised framework, raw features are rarely informative and need a baseline. Speaker normalization is expected to automatically incorporate corpus channel variability, with reduced accuracy for larger channel variance. A speaker's baseline (neutral) is modeled as a single Gaussian per feature computed using all neutral-labeled utterances. In the case of the VAM database, we performed multiple experiments, one using normalization regardless of arousal label and one in which a limited number of the earliest neutral arousal labels was used for normalization.

3.3. Arousal Rating Computational Framework

In supervised learning, model parameters are learned based on a set of labeled data. Instead of training a high-dimensional model on one corpus and testing it on another, we approach the problem from another angle. Given a baseline for a speaker, we seek to fuse multiple ratings from diverse knowledge-based features to get a final arousal rating. This adaptive, summative framework is analogous to human information processing– it has been demonstrated that sensory integration, even within a single mode, is linear with dynamic weights assigned by apparent cue reliability [19].

First we obtain neutral models for each speaker. Then, feature x_i of utterance i, $i\epsilon\{1, 2, 3\}$, is scored, p_i, on the corresponding neutral Gaussian, N_i, as

$$p_i = 2 \times cdf_{Ni}(x_i) - 1$$

where $cdf_{Ni}(\cdot)$ is the cumulative distribution function of single-variate Gaussian N_i. The score is bounded in the range $[-1, 1]$. This score is computed on the neutral class as well in the case of IEMOCAP, for which we have both categorical and dimensional labels.

Ratings are fused by normalized weighted summation. The weights are calculated per-speaker as the Spearman's rank-correlation of each score vector $\mathbf{p}_i$ with the score mean vector $\mathbf{p}_\mu$, inspired by Grimm [10].

4. Results and Discussion

Our primary goal is to generate a continuous rating that correlates with human annotations of arousal. However, in order to generate comparable results for future studies, binary classification is also performed. We arbitrarily set the threshold to 0 on our arousal ratings to classify high and low arousal. Unweighed average recall, the average of the recalls of both classes, is reported.

4.1. Acted Speech Arousal Rating

The results for all three acted emotional databases (Table 2) demonstrate that our arousal rating framework provides an interpretable measure of arousal, given a minimal set of assumptions. It is important to note that intensity and pitch are the most highly correlated features with arousal labels for different databases. Fusion often provides higher correlations than any single feature.

We obtain significant correlations of 0.77 and 0.70 for EMA and emoDB, respectively. Weighted fusion correlation exceeds that of unweighted fusion in both cases. The emoDB intensity-subrating had an undesirably high negative correlation with arousal labels (although we knew a priori that the intensity information is unreliable). It is desirable to place very little weight on intensity in this instance, and our fusion framework accomplishes this task by assigning a weight of -0.02. Weighted fusion increases correlation between the arousal rating and arousal labels from 0.62 (UW fusion) to 0.70. Histograms of arousal ratings show clear biases of the four emotional categories to high and low arousal (Figure 1).

Arousal ratings generated on IEMOCAP correlate well, ρ=0.62, with the mean rater labels– they are plotted as a surface in Figure 2. Results were also generated separately for the improvisation and scripted portions of the IEMOCAP database, but no notable difference exists.

When we perform unsupervised binary classification, we obtain unweighted average recalls of 86%, 82%, and 73% on EMA, emoDB, and IEMOCAP, respectively. Recalls are well above chance and results on emoDB compare well to those presented by Schuller et al. [9].

Furthermore, when scoring with only 5 neutral utterances per speaker in IEMOCAP (compared to $\sim$100), we achieve similar results of ρ =0.58 and a recall of 72%, demonstrating that little data is needed for normalization.

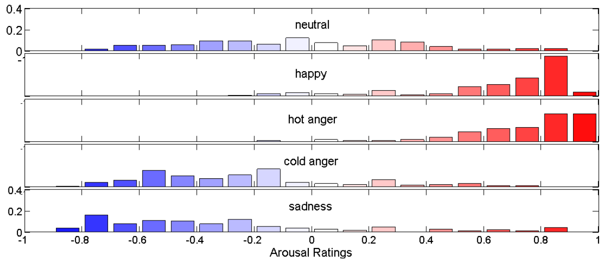

Figure 1: *Histograms of EMA Arousal Ratings for categorical emotions. Sp.* $\rho = 0.77$ *(*$p < 10e{-}100$*).*

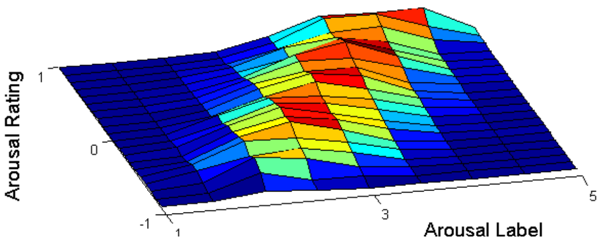

Figure 2: *Histogram of IEMOCAP results. Sp.* $\rho = 0.62$ *(*$p < 10e{-}100$*). Hotter colors indicate higher occurence.*

4.2. Natural Speech Arousal Rating

Correlation coefficients are reported in Table 2 for the VAM database. Normalization is performed in two ways. In one experiment, we consider only speakers with at least 10 utterances, choosing up to four that are closest to 0-rated arousal for neutral modeling. In this limited case, only 36 of 47 speakers and 870 of 941 sentences were investigated. In the second experiment no requirement is placed on speakers, and we perform speaker normalization regardless of arousal labels.

Results demonstrate that the algorithm is effective in the given natural emotional speech database. Classification accuracy in the limited neutral speaker-normalization case is 73%, slightly higher than the 70% in the general case– which did not require explicit neutral state labeling. While the overall correlations are medium-high at 0.65 and 0.63, within-speaker correlations are in the range [-0.05,0.93] for the speaker-baseline enrollment trial. Although correlations on as few as 10 utterances should be interpreted with caution, other explanations may be that some speakers had limited emotional variability or had few training samples near their true baseline.

The results compare favorably to another approach which uses many features, while making no assumptions or normalizations across corpora or speakers [8]. We achieve 17% absolute improvement over that result. Our framework is simple and makes the primary, but practical assumption that a small sample of neutral data is known.

5. Conclusions

We have introduced an unsupervised approach as an alternative to supervised, cross-corpora classification for the rating of arousal. The framework makes the assumption that labeled neutral speaker data is available, although some robustness to this assumption may be interpreted from our results on a natural, spontaneous emotional speech corpus and that only a few samples are needed on IEMOCAP for peak performance. In practice, neutral data can be collected offline, or in an interview setting.

Several directions for future work are possible. Additional features may be chosen depending on the domain such as lexical or voice quality features like NAQ. Results using the same framework for more rating tasks should be investigated. It is interesting to consider attempting valence rating, but it is well known that obtaining valence from speech features is much more difficult than arousal. Future work will also include application to Behavioral Signal Processing (BSP) domains such as couple therapy and autism [20].

6. Acknowledgements

This work was supported by funds from NSF and NIH.

7. References

[1] B. Z. Pollermann, “A Place for Prosody in a Unified Model of Cognition and Emotion,” in *Proc. Speech Prosody*, 2002.

[2] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. la Torre, “Detecting Depression from Facial Actions and Vocal Prosody,” in *Affective Computing and Intelligent Interaction*, September 2009.

[3] M. Nowak, J. Kim, N. W. Kim, and C. Nass, “Social Visualization and Negotiation: Effects of Feedback Configuration and Status,” in *Proceedings of ACM CSCW*, 2012.

[4] P. Juslin and K. Scherer, *The New Handbook of Methods in Nonverbal Behavior Research*. Oxford: Oxford University Press., 2005, ch. 3. Vocal Expression of Affect, pp. 65–135.

[5] C. M. Lee and S. Narayanan, “Towards detecting emotions in spoken dialogs,” *IEEE TASLP*, vol. 13, no. 2, pp. 293–302, 2005.

[6] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Communication*, vol. 53, pp. 1062–1087, 2011.

[7] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, “Cross-corpus Classification of Realistic Emotions- Some Pilot Experiments,” Proc. of Inter. Workshop on EMOTION 2010, pp. 77-82.

[8] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, “Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote?” in *Proceedings of Interspeech*, 2011.

[9] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, “Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies,” *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[10] M. Grimm and K. Kroschel, “Evaluation of Natural Emotions Using Self-Assessment Manikins,” in *Proc. of IEEE ASRU*, 2005.

[11] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database,” *J. of Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[12] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, “An Articulatory Study Of Emotional Speech Production,” in *In Proc. Eurospeech*, 2005.

[13] J. Kim, S. Lee, and S. S. Narayanan, “An Exploratory Study of the Relations between Perceived Emotion Strength and Articulatory Kinematics,” in *Proceedings of Interspeech*, 2011.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A Database of German Emotional Speech,” in *Proceedings of Interspeech*, 2005.

[15] M. Grimm, K. Kroschel, and S. Narayanan, “The Vera am Mittag German Audio-Visual Emotional Speech Database,” in *Proc. of IEEE ICME*, 2008.

[16] K. Scherer, “Vocal Affect Expression: A Review and a Model for Future Research,” *Psychological Bulletin*, vol. 99, 1986.

[17] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[18] M. K. Sonmez, L. Heck, M. Weintraub, E. Shriberg, M. Kemal, S. Larry, H. Mitchel, and W. E. Shriberg, “A Lognormal Tied Mixture Model Of Pitch For Prosody-Based Speaker Recognition,” 1997.

[19] M. Young, M. Landy, and L. Maloney, “A Perturbation Analysis of Depth Perception from Combinations of Texture and Motion Cues,” *Vision Research*, vol. 33, pp. 2685–2696, 1993.

[20] M. P. Black, D. Bone, M. E. Williams, P. Gorrindo, P. Levitt, and S. S. Narayanan, “The USC CARE Corpus: Child-Psychologist Interactions of Children with Autism Spectrum Disorders,” in *Proceedings of Interspeech*, 2011.