Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features

Daniel Bone, *Senior Member, IEEE*, Chi-Chun Lee, *Member, IEEE*, and
Shrikanth Narayanan, *Fellow, IEEE*

Abstract—Studies in classifying affect from vocal cues have produced exceptional within-corpus results, especially for arousal (activation or stress); yet cross-corpora affect recognition has only recently garnered attention. An essential requirement of many behavioral studies is affect scoring that generalizes across different social contexts and data conditions. We present a robust, unsupervised (rule-based) method for providing a scale-continuous, bounded arousal rating operating on the vocal signal. The method incorporates just three knowledge-inspired features chosen based on empirical and theoretical evidence. It constructs a speaker's baseline model for each feature separately, and then computes single-feature arousal scores. Lastly, it advantageously fuses the single-feature arousal scores into a final rating without knowledge of the true affect. The baseline data is preferably labeled as neutral, but some initial evidence is provided to suggest that no labeled data is required in certain cases. The proposed method is compared to a state-of-the-art supervised technique which employs a high-dimensional feature set. The proposed framework achieves highly-competitive performance with additional benefits. The measure is interpretable, scale-continuous as opposed to discrete, and can operate without any affective labeling. An accompanying Matlab tool is made available with the paper.

Index Terms—Arousal, activation, rule-based rating, knowledge-inspired features, cross-corpora classification, continuous affect tracking

1 Introduction

Emotion is at the core of human behavior, influencing our decisions both consciously and unconsciously. A primary case is communication, which is the process of exchanging information between sender and receiver to achieve a "commonness" of interpretation [1]. Human communication is an intricate process in which the participants are constantly transmitting information through verbal and non-verbal (vocal, visual, gestural) cues, even when not the active speaker. During many interactions, the participants will display overt affective signals that are invaluable to moderating the tone of the exchange. As examples, consider the importance of affective cues for: an employee negotiating a raise in salary with an employer; teasing amongst friends while being cognizant of taking it too far; a spouse who wants to know when it is imperative to act on their partner's requests. In all of these cases, accurate interpretation of a person's tone of voice or facial expression may keep a positive rapport or help achieve a desired outcome.

Affective phenomena, especially arousal, inspire scholarly interest in many disciplines for diverse purposes. These disciplines include psychology and sociology, biology, engineering, linguistics, and even consumer research. For instance, human behavior studies in adults have investigated concepts such as interpersonal intimacy (e.g., interpersonal distance, eye contact, and touch); one model proposes that one person's acts of intimacy predict an arousal change in the other person [2]. Other work has linked activation (also referred to as arousal) to differences in task performance; thus activation is considered a measure of personal motivation [3]. Affect study in children is also of significant interest [4]; for example, one experiment investigated the arousal of infants in relation to depression in their mothers [5]. Furthermore, receptive and expressive emotional processing are critical to the understanding of the prevalent neurodevelopmental social disorder autism; studies have examined non-verbal communicative performance [6] and emotional responses in the brain [7]. It is apparent that applications of affective computing technologies are abundant.

Humans can judge affective content from voice at accuracies well-above chance, and speech processing techniques can do so as well. However, a common finding among speech acoustic studies is that non-specific vocal arousal is identified more effectively than pleasantness (valence) [8]; this is also shown empirically in cross-corpora classification studies [9], [10]. Therefore, we suggest that vocal arousal is an area primed for creating general-use tools.

Pollerman (2002) has posited that prosody is an essential mechanism for investigating cognition and emotion [11]. In many cases prosodic correlates of arousal are used as variates for analyzing human behavior. Given the importance of arousal, it also is understandable that researchers need various (sometimes simultaneous) measures. However, there is an absence of available validated tools for measuring arousal from voice. This is the primary motivation for this work and for the accompanying vocal arousal rating

- *D. Bone and S. Narayanan are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA, 90089. E-mail: dbone@usc.edu, shri@sipi.usc.edu.*
- *C.-C. Lee is with the Electrical Engineering Department, National Tsing Hua University, Taiwan. E-mail: cclee@ee.nthu.edu.tw.*

Manuscript received 15 Nov. 2013; revised 10 May 2014; accepted 14 May 2014. Date of publication 29 May 2014; date of current version 23 July 2014.
Recommended for acceptance by B.W. Schuller.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2014.2326393

tool. In this section we describe standard arousal measures in behavioral science research, the state of affective classification, the necessity of a convergence between disciplines, and the study goals and design.

1.1 Affect in Behavior Sciences

A person's internal affective state can be measured through various methods, both qualitative and objective. Many times researchers must rely on a qualitative measure such as patient self-report. For example, the Positive and Negative Affect Scale (PANAS) is a prevalent 10-item questionnaire that produces varimax rotations of arousal and pleasantness (valence) [12]. Another popular self-report measure is the behavioral inhibition scale (BIS) or behavioral activation (BAS) scales [13]. Objective measures do not rely on rater bias, although they can be influenced by other factors. Objective measures of autonomic arousal include heart rate, electro-dermal activity (EDA), and pupil diameter. For instance, pupil diameter has been linked to emotional arousal associated with increased sympathetic activity [14]. But another overt behavioral signal is available "for free" in many datasets.

Researchers are increasingly utilizing the human voice as a non-invasive measure of expressed arousal. Audio is available in many situations where other objective measures are not an option. Audio can even be useful in situations where the collected data were not originally intended for arousal research. Furthermore, a researcher may be interested in having multiple measures of arousal for increased reliability or to study relationships between different arousal constructs. For instance, certain vocal cues can be modified such that they do not reflect the speaker's true internal state in order to persuade or dissuade a listener [8]; having separate measures of overt and covert affect could potentially reveal this. But currently the availability of arousal rating tools is limited to non-existent. Many behavioral studies currently rely on an individual prosodic feature as the sole measure of expressed arousal. For example, pitch has been used to investigate outcomes of depressed patient interventions [15] and to produce a visual depiction of arousal for negotiators [16].

It is intuitive that expressed vocal arousal will modulate more than a single quantified feature, and thus this will be a sub-optimal arousal tracker. An individual feature may not have consistent meaning in different social contexts or audio settings, but the inclusion of more features in a vocal arousal measure could provide robustness. For example, Pollermann (2002) created a vocal arousal measure by combining z-normalized vocal pitch, intensity, and speaking rate (SR) for use in two interesting pilot studies [11]; one study compared autonomic arousal in patients with autonomic lesions to the patients' expressed (vocal) arousal, while another distinguished the adaptive coping abilities of patients with breast cancer based on their vocal arousal. In general, this feature combination strategy has not been fully validated for effectiveness in capturing arousal. Producing a validated measure of vocal arousal is a capability of affective computing.

1.2 Affective Computing and Vocal Arousal

There is an impressive history in engineering of investigating affect from voice [17], [18], [19], [20], [21]. Engineers have primarily investigated emotion for its utility in systems applications. Affective computing has shown potential for the development of intelligent human-machine interfaces [20] and for improving core speech technologies like speech recognition and speaker identification [22]. However, due to limitations in available data and the relative infancy of the field, much of the research has focused on maximizing predictive performance within a single dataset (e.g., the Speech Under Simulated and Actual Stress database [23]). The success of affect recognition is highly-dependent on the characteristic of a specific corpora and the applied computational techniques. Correspondingly, there has been little agreement on the features that are predictive of affective constructs.

Consequently, cross-corpus robustness of emotion recognition is gaining interest. This approach aims to build a system that is capable of handling diverse acoustic settings and speaker traits (e.g., gender and age), while maintaining high performance. Eyben et al. [24] and Schuller et al. [9] have pursued cross-corpora dimensional emotion classification. The authors developed systems that did not require any data for corpus-adaptation (i.e., corpus or speaker normalization), achieving above-chance accuracies ($\approx$60%). Still, the accuracies were lower than desired for many applications and Eyben et al. (2010) noted the need for methodology to mature. One approach to improving results is speaker- or corpus-normalization, in order to adapt the system to new attributes. Schuller et al. (2010) achieved much higher accuracies across corpora through speaker normalization [10]. Such efforts in robust emotion recognition may be applicable not only in engineering systems, but in creating affective ratings for behavioral studies.

1.3 An Opportunity for Convergence

Human behavior researchers are in need of robust measures of affective constructs that are transferrable across corpora; this is a great opportunity for engineers to employ methodologies to create simple veritable measures of emotion. More specifically, engineers need to first consider the goals of an application domain, and then design the required system. In our case, engineering can provide an arousal rating tool which: incorporates more features than the standard *mean pitch* used by many psychologists as a measure of arousal; generalizes well (possibly better than high-dimensional feature vectors that are susceptible to over-fitting); does not require labeled training data; maintains interpretability; and is simple to use.

More detailed points reflecting our views for this convergence follow. First, it is understandable that *mean pitch* may be insufficient as a measure of vocal arousal, since a speaker may display emotions through other cues and modalities. Juslin & Scherer (2005) state that interactions between features may reflect combinations of measures more closely related to human perception; for instance, 'vocal effort' may be a combination of acoustic features such as vocal intensity and high-frequency energy [25]. Thus, integrating multiple variables can lead to better modeling and potentially increased robustness. Yet, pitch as a measure of vocal arousal has the benefit of maintaining interpretability of the model.

Second, engineers must create algorithms that generalize well, but also accommodate the constraints of the target domain. A supervised approach has many challenges. For instance, data in the suggested application domains will often not contain any associated labeled data that would be useful for model adaptation. Also, since emotion classification from speech is strongly influenced by phonetic structure [26], a supervised system risks being dependent on the phonetic structure of the data on which it was trained. In our specific case, we find that providing computational robustness to the models similar to those already employed in psychology is a suitable approach. Lastly, fundamental signal processing techniques like speaker normalization can benefit cross-corpus affect modeling; such techniques are not universally applied in behavioral research, even when necessary. In our preliminary work, we developed an arousal rating framework that addresses the previously stated objectives of accuracy, robustness, and interpretability [27].

1.4 Study Goals and Design

In this work, we aim to develop and validate an engineering framework for vocal arousal rating that adheres to the constraints of the target domain, behavioral science. Our proposed system is simple, incorporating only three acoustic features and not requiring labeled emotional training data. Our system is also robust, achieving high correlation and classification accuracy in diverse scenarios; we evaluate multiple languages (German and English), emotional contexts (scripted and read), and emotional styles (acted and natural). This framework is also generally-applicable if known, robust correlates of a target variable exist.

A brief overview of our unsupervised (rule-based), knowledge-inspired system follows. The chosen features are knowledge-inspired, based on the survey article by Juslin & Scherer (2005) which defines acoustic correlates of vocal arousal that have consistently predictable behavior across many empirical studies [25]. Moreover, these empirical results are also predicted by anatomical models of affective speech production; e.g., pitch is expected to increase when stress or arousal causes the muscles in the larynx to tighten. In this work, we investigate an array of features indicated by perception and production evidence. Our final selected feature set contains median log-pitch, median intensity, and HF500 (similar to spectral slope); this is identical to the feature set selected by intuition and feature extraction constraints in our previous work [27]. This small, interpretable feature set can be robustly extracted and leads to coherent results across corpora.

With our chosen feature set, we obtain an soft-rating of arousal for each feature. We do not assume a Gaussian distribution of a speaker's features as in our previous work [27], but use the exact values of the baseline data as an estimate of the distribution. The only requirement of the algorithm is that neutral data is available to define a speaker's neutral feature range (since it is well-known that a speaker's features are idiosyncratic [8]); however, we will demonstrate the algorithm's utility even if no such labeled neutral data is available. Finally, the soft-ratings from each feature are combined through weighted summation using a method that does not rely on labeled training data, but the relation between soft-ratings for a given speaker and corpus [28]. This provides demonstrated robustness when one feature is corrupted. In this work, we additionally compare our system to a supervised baseline and explore temporally-continuous arousal ratings that do not rely on utterance boundaries.

We describe the system as *unsupervised* since feature weights (model parameters) are not adjusted such that the system output matches the labels in a set of training data, as with supervised approaches. As such, we do not perform *cross-corpora classification* in which we train model parameters on one or more databases and test the model on a unique database; instead, we perform *cross-corpora evaluation* of our rule-based system. The system essentially performs a type of speaker-normalization using the speaker-baseline model, but this requirement is of little consequence since supervised cross-corpora approaches have had little success without speaker normalization.

In the Methods section, we detail the emotional corpora and our proposed arousal rating framework. In the next section, we detail our results and the corresponding experimental setup; this includes validation of our approach across multiple corpora, testing of other potential features such as speech rate, and evaluation of alternative supervised techniques with state-of-the-art features and various adaptation strategies. In the remaining sections we discuss results and potential applications of this arousal rating tool, and then conclude.

2 Methods

Our experiments are conducted with five emotional databases comprising acted and natural German and English speech with dimensional and categorical arousal labels. The databases (detailed in Table 1) are: IEMOCAP, emoDB, EMA, VAM, and CreativeIT. The first four are publicly available presently. CreativeIT includes temporally-continuous arousal ratings.

2.1 Databases

2.1.1 Acted Emotional Speech Corpora

IEMOCAP consists of mixed-gender dyadic interaction between actors speaking in English, along with associated categorical and dimensional emotional labeling [29]. Five dyads interact in both spontaneous improvisation of hypothetical emotional scenarios (2,388 turns) and portrayal of scripted emotional content (4,517 turns). At least two raters rate every turn. Our analyses concentrate on the dimensional arousal rating, which is a five-pt scale. Neutral turns were defined by the categorical label "neutral". The data initially contains 10,039 turns. A turn was excluded if speech was overlapped by another speaker or there was significant background noise (3,134 turns). In addition, some utterances had no voiced frames identified by Praat ($<1\%$), and were discarded. We include the data for which no agreement was made regarding categorical emotion (28 percent).

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. Four raters labeled the categorical emotion of each utterance. Categorical labels of hot anger and happy are denoted as high arousal in our study, while sadness and cold anger

TABLE 1
Description of Emotional Corpora and Arousal Labels

Corpus	Style	Emotion	Label	–	+	Neu	Total	Speakers	Setting	Language
IEMOCAP	spontaneous & scripted	acted	ordinal	2,579	4,304	(1,112)	6,883	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
CreativeIT	induced	acted	continuous	-	-	N/A	-	16 (8f,8m)	studio	English

are designated low arousal. The speech was intentionally modulated with different speaking styles: normal, loud, and fast; these variations may affect arousal perception, especially as attributed to energy and speech rate. The speakers articulation may also be somewhat hindered by the sensors placed on the face and tongue for the purpose of studying affective articulation with electromagnetic articulography [30], [31].

emoDB is comprised of acted (read) German speech. Seven emotions are expressed: neutral (neutral arousal); happy, angry, and fearful (high arousal); and sad, bored, and disgusted (low arousal). The acoustic intensity is unreliable due to varying mouth-to-mic distance [32]; thus, a system that places heavy weight on vocal intensity would fail. We will evaluate our score-fusion scheme in later sections for the task of addressing this corrupted feature.

Our fourth acted emotional corpus is CreativeIT, which incorporates the Active Analysis improvisation technique to encourage goal-oriented affective interactions [33]. Time-continuous (downsampled to 100 Hz) and scale-continuous (in the range $[-1, 1]$) arousal annotations were made by two or more raters; the ratings were variance-normalized and averaged. The processed-data we incorporate contains 16 actors who participated in 45 dyadic interactions; thus, we have 90 total tokens for temporally-continuous analysis of objective vocal arousal in relation to subjective perceived arousal. Since the absolute value of the ratings isn't necessarily important, we do not require neutral labels in this database and opt for *all-data normalization* (Section 2.4).

2.1.2 Natural Emotional Speech Corpus

The VAM corpus [34] consists of speakers in dyadic or triadic conversations on the German TV talk-show "Vera am Mittag", or Vera at Noon. The data is spontaneous and considered natural, although the naturalness of talk-show speech is debatable. Forty-seven speakers account for a total of 947 utterances. Some of the speakers speak less than 10 utterances and are disregarded from these analyses; 36 speakers and 870 utterances are investigated. The utterances are labeled for arousal (also valence and dominance) on a continuous scale by 7 to 16 raters. Since no explicit neutral tag is given for the data, we select up to four utterances that have closest to 0-rated arousal for baseline modeling.

2.2 Knowledge-Inspired Features

The foremost principle in the design of our rule-based system is that certain features display predictable trends across many contexts. As mentioned previously, mean-pitch is commonly taken as a measure of vocal arousal, since robust cross-corpora arousal rating systems are not yet freely-available. In order to construct a robust arousal rating system, we incorporate solely features that are indicated through both perception and production research. These features are consistently predictive of perceived arousal in many empirical experiments, neatly summarized by Juslin & Scherer (2005, in Table 3.2). The empirical results are not surprising, as the feature trends can also be predicted by both discrete emotion theories and component process theory [25]. For instance, fear (high arousal) causes the laryngeal folds to tighten as a sympathetic response, leading to higher pitch [35].

Our review indicates that five features are regularly reported to be affected by increased arousal: pitch (mean and variance increase); vocal intensity (mean and variance increase); HF500, a voice quality measure which is the ratio of high-frequency to low-frequency energy with a 500 Hz cutoff (increases); speaking rate (increases); and jitter, a measure of pitch aperiodicity (increases). However, some uncertainty remains about the contextual dependence of these features and the robustness of automatic extraction. Based on our own analysis, we incorporate three features into our final model. The final knowledge-inspired features that we incorporate are median pitch, median vocal intensity, and HF500. Median is used for its resilience to outliers. It is important to note that all features are extracted only on voiced frames (determined by Praat). HF500 is more specifically computed as the ratio of amount of energy above 500 Hz to the amount of energy between 80 and 500 Hz, removing low-frequency noise. Pitch, vocal intensity, and jitter are extracted using Praat with a 25 ms window and 10 ms shift. Speaking rate is calculated using manual annotations of word boundaries, as well as with an automatic energy-based method which depends on sensitive thresholds [36].

2.3 Rule-Based Arousal Rating Framework

Typically, model parameters are learned through supervised approaches that incorporate sets of labeled data.

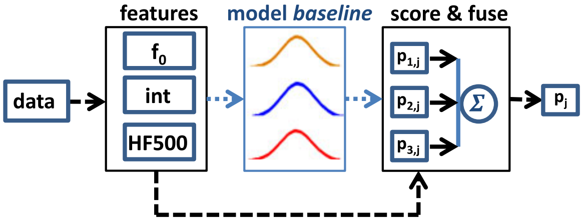


Fig. 1. Arousal rating flow diagram. Utterance j is first transformed into three feature streams. Each feature produces an arousal score based on a baseline model trained for each individual. Lastly, the scores are combined into a final arousal rating, p_j.

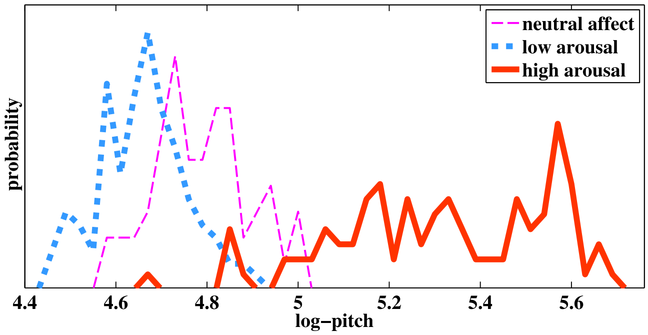


Fig. 2. Probability density functions of log-pitch for one speaker from the EMA database. Labeled *high arousal* tends to indicate increases in pitch over neutral.

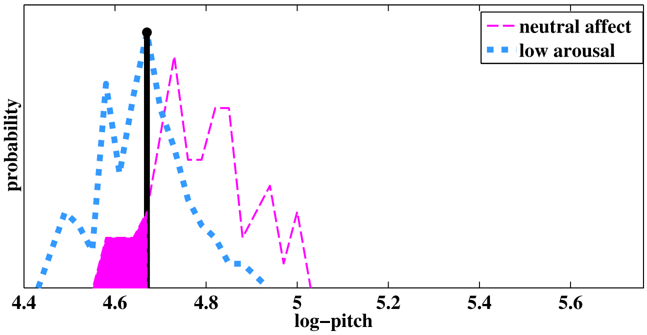


Fig. 3. Arousal rating of a token with labeled *low arousal* having a log-pitch of 4.67. The filled-in portion represents the fraction of neutral samples which are lower than 4.67.

When training with high-dimensional feature vectors but hoping for cross-corpora robustness, overfitting to emotional contexts, phonetic structure, and speaker characteristics of a training database is a genuine concern. Researchers have trained on multiple corpora in an effort to model multiple contexts [9], [10], [24]. We approach the problem from a different angle based on the consistent predictability of certain acoustic correlates of arousal in relation to a reference point. Given a baseline for a speaker, we fuse multiple ratings from these diverse knowledge-inspired features to produce a robust rating of vocal arousal in the range $[-1, 1]$. This summative, adaptive (the fusion framework is adaptive) technique may have similarities to human perceptual processes—i.e., sensory integration, even within a single mode, has been shown to be linear with dynamic weights assigned according to cue reliability [37].

Our framework (depicted in Fig. 1) begins by extracting select knowledge-inspired features from the acoustic waveform. Baseline (neutral) behavior of each feature is modeled per speaker. Arousal scores are generated for each feature based on expected trend with arousal and the baseline model. A final arousal rating is computed through weighted summation of single feature scores. The arousal rating framework is further detailed in the remainder of this section.

2.3.1 Rule-Based Decisions

One approach to creating an unsupervised classification system is the inclusion of pre-defined rules. We predefine the direction of the relationship between changes in the knowledge-inspired features and changes in vocal arousal (Table 2). In particular, we adopted the rules that increases (decreases) in pitch, intensity, and HF500 are indicative of increases (decreases) in vocal arousal. Having set these rules, the system still needs to understand if a certain feature value represents an increase or decrease in arousal, and thus must have a model of a speaker's baseline. Additionally, a fusion strategy should be incorporated to combine potentially conflicting scores from the individual features.

TABLE 2
Chosen Rules For Arousal Rating Framework, Base on Knowledge From Juslin & Scherer (2005) [25]

	Median Pitch	Median HF500	Median Voc. Int.
Expected/Defined Change for Increased Arousal	↑	↑	↑

2.3.2 Speaker Baseline Modeling

Raw feature values are rarely informative without a reference; often the variability between speakers is larger than variability within speakers. The proposed framework depends on a baseline to assess deviations in feature values, and thus deviations in vocal arousal. Our neutral (baseline) model is simply a vector containing feature values from all neutral (baseline) tokens. When making decisions per utterance, we assign the median feature value of all the voiced frames within an utterance as the utterance's feature value. If providing ratings at every voiced frame in an audio file, all such frames are used for baseline modeling. Since each feature will have a different baseline, a separate model is created for each feature. This baseline model is represented as a probability density function (pdf) in Fig. 2. Speaker normalization is expected to account for corpus channel variability (as shown by Schuller et al. (2010) [10]).

In cases where labeled neutral data is not available, all the data for a particular speaker may be used in the baseline model. The effect of the type and amount of baseline data is investigated in Section 3.3. The drawback of not having labeled neutral data is that arousal rating values should become more relative. In particular, a negative rating may no longer suggest that the arousal is negative, only that it is negative in comparison to the baseline data.

2.3.3 Arousal Scoring and Fusion

Having chosen our set of three features and established the need for a baseline model, we will detail our generally-applicable framework for scoring and fusion. We first acquire neutral (baseline) models N_i for each feature type $i\epsilon\{1, 2, 3\}$ as described in Section 2.3.2. Then, feature value $x_{i,j}$ of utterance j, is given a score, $p_{i,j}$, with the corresponding neutral model, N_i, by

$$p_{i,j} = 2 \times E[x_{i,j} > N_i] - 1,$$

where $E[x_{i,j} > N_i]$ is the percentage of neutral model (N_i) values for which $x_{i,j}$ is larger. The score is bounded in the range $[-1, 1]$. If the baseline model was given neutral labels, the scores may be interpreted such that positive (negative) scores indicate positive (negative) arousal, with magnitude being associated with confidence. The framework is simple mathematically, but has the advantage of being very interpretable. An example is shown in Figs. 2 and 3.

In some databases, one of these features may be corrupted, as is the case in the emoDB corpus (Section 2.1.1). A smart fusion strategy will be able to ignore that feature; however, we cannot use any arousal labels in this fusion strategy in order to make this framework be available for many applications. In order to combine the scores from each feature into a single arousal rating, we employ a technique which requires no arousal labels (inspired by Grimm et al. (2005) [28]). The weights for fusion are calculated per-speaker as the Spearman's rank-correlation coefficient between each score vector $\mathbf{p}_i$ and the score mean vector $\mathbf{p}_\mu$, where the vectors are composed of scores for all of a speaker's utterances. Weights are then normalized to have a combined magnitude of 1.

TABLE 3
Feature Comparison for Arousal Scoring in Reference to Spearman's Rank Correlation, ρ_S, with Arousal Labels

Feature		Corpus			
Type	**Functional**	*IEMOCAP*	*emoDB*	*EMA*	*VAM*
log-pitch	*median*	**0.50**	**0.75**	**0.54**	**0.56**
	IQR	0.32	0.14	0.31	0.32
	range	0.28	0.13	0.31	0.32
	floor	0.37	0.71	0.39	0.31
intensity	*median*	**0.59**	**−0.46**	**0.67**	**0.65**
	IQR	0.23	0.34	0.44	−0.16
loudness	*median*	**0.61**	**0.07**	**0.68**	**0.64**
HF500	*median*	**0.45**	**0.69**	**0.65**	**0.37**
jitter	*median*	−0.15	−0.21	−0.02	0.05
HNR	*median*	0.18	−0.34	0.02	−0.13
shimmer	*median*	−0.06	0.19	−0.02	0.09
true SR	*median*	−0.06	-	−0.03	-
est. SR	*median*	0.08	0.23	0.18	−0.05

2.4 Supervised Arousal Classification Baseline

In order to compare our result to a more exact baseline, we emulate the state-of-the-art approach of using openSMILE features [38] with linear SVM. This approach sets a formidable baseline for many Interspeech Challenges (e.g., [39], [40], [41]). We used a configuration file from the 2011 Interspeech Challenge which extracted 4,368 features that cover the spectrum of commonly utilized spectral and prosodic descriptors. Linear SVM models (L2-loss and L2-regularization) were built using LIBLINEAR [42]. Separate models were trained on each corpus individually, and then tested on the remaining corpora; a majority-vote decision was made for each test corpus (based on the other three corpora). The cost parameter was optimized for each training database in the set $10^{[-4,-3,-2,-1,0,1]}$.

Our proposed framework generally uses all neutral data as a baseline for each speaker. We employ three normalization strategies for the supervised approach: *no normalization*, *neutral-data normalization*, and *all-data normalization*. Speaker z-normalization (mean and variance normalization) is used because it has proven successful in similar tasks [10], [43]. In order to control for channel (e.g., mouth-to-mic distance) and speaker (e.g., vocal morphology) properties, normalization is necessary; thus, *no normalization* does not account for these differences. With *neutral-data normalization*, baseline parameters for each speaker are trained using only neutral data for that speaker (as is the case in our proposed model). *All-data normalization* trains baseline parameters on all of a speaker's data. This method benefits from increased data for baseline modeling and can handle unlabeled data. However, *all-data normalization* is influenced by the distribution of arousal for each speaker; in particular, a decline in classification accuracy is expected if training on data with a majority of high-arousal instances and testing on data with a majority of low-arousal instances.

3 Experimental Setup and Results

Our primary goal is to create and validate a versatile arousal rating tool that measures expressed arousal. Since a ground-truth measure of internal arousal is not available, we correlate with and predict human annotations of perceived vocal arousal. In the first experiment, relations between knowledge-inspired features and arousal labels are assessed, leading to the selection of three features for inclusion in our final model. We consider more features than the three chosen for our initial work [27]. Second, we compare our selected features and framework to state-of-the-art supervised approaches for cross-corpora binary (high/low) arousal recognition [38]. Third, we examine the amount of neutral data required in terms of system accuracy, as well as an approach to arousal rating without any neutral-labeled data. In our final experiment, we demonstrate a modified approach for continuous arousal rating with temporal smoothing in the CreativeIT database.

3.1 Acoustic Feature Comparison

Arousal scores are generated for various knowledge-inspired features in order to select those that are the most robust for our framework within these datasets. We consider features from the following categories: pitch, volume, voice quality, and speaking rate. The correlations between the arousal score from each feature (before any fusion) and the manually-labeled arousal are displayed in Table 3. In these experiments, we include the scores for the neutral-designated data in the correlations, which reduces final correlations since neutral data will uniformly receive scores in the range $[-1, 1]$ per speaker. This section considers only utterance-level decisions.

Four utterance-level functionals of log-pitch are considered: median, inter-quartile range (IQR), range, and floor. Median and IQR are robust analogues of mean and standard deviation. Robust floor (10 percent quantile) and robust range—the difference between the robust ceiling (90 percent quantile) and robust floor (10 percent quantile) in an utterance—are included because they were informative in many psychological studies. The results clearly indicate that modeling median of log-pitch produces a strong correlate of arousal.

Results for three volume features are reported: median and IQR of vocal intensity (short-time energy), and median

TABLE 4
Median Spearman's Correlation Coefficient between Selected Features in Utterance-Level Emotional Corpora

	Corpus			
Feature	*IEMOCAP*	*emoDB*	*EMA*	*VAM*
log-pitch & HF500	0.41	0.78	0.53	0.21^{ns}
log-pitch & vocal intns.	0.61	−0.49	0.59	0.54^{*}
HF500 & vocal intns.	0.66	−0.69	0.78	0.30^{ns}

Sig: ns- $p >= 0.05$*; ∗-* $p < 0.05$*; else-* $p < 1e-4$.

loudness.[1] Loudness computation is motivated by the perceptual work of Zwicker et al. (1991) [45]. Vocal intensity is computed through Praat. IQR of vocal intensity produces the lowest correlations. Median vocal intensity and loudness deliver comparable correlations, except for the emoDB corpus; the median-vocal-intensity score has an unexpected medium-strength negative correlation, while loudness produces a very small positive correlation as expected. However, the feature weighting framework compensates for this issue by assigning near-zero weight to vocal intensity (Section 3.2). Since loudness computation has much higher computational cost, median of vocal intensity is selected from this group for the final arousal rating framework.

Four voice quality features are experimented with: HF500, jitter, HNR (harmonics-to-noise ratio), and shimmer. HF500 is the ratio of high-frequency to low-frequency energy with a 500 Hz cutoff. Jitter is a short-term pitch aperiodicity measure, while HNR is a measure of acoustic periodicity; they are typically negatively correlated. Shimmer is a measure of short-term amplitude perturbation. Jitter, HNR, and shimmer are calculated in Praat. HF500 is calculated from the long-term average spectrum using a script written in Matlab. Although jitter was indicated to be an acoustic correlate of arousal [25], the only consistent correlate of arousal is HF500.

Speaking rate is a common correlate of vocal arousal [25]; but it is dependent upon many potentially confounding factors. We consider two methods of obtaining speaking rate. With both of the methods we use the median syllabic speaking rate (syl/s) of all syllables in the utterance as the measure of SR for an utterance. The first method, called *true SR*, uses syllable boundaries computed from phonetic forced-alignment of speech to text (available only for EMA and IEMOCAP). The second method, called *est. SR*, estimates the syllable boundaries using a lexically-independent, energy-based syllable-nuclei detection method [36]. The true and estimated SR have medium correlation in the EMA corpus $\rho_S = 0.62$ (p < 1e-10) and low correlation in the IEMOCAP corpus $\rho_S = 0.18$ (p < 1e-10). The results suggest that neither true or estimated speaking rate are consistently informative of vocal arousal for these corpora and our modeling approach.

Ideally, the final feature set will not only show strong predictive potential, but also the features should be diverse enough to profit from fusion. For instance, the features should not be affected by channel and speaker variations or distortions in the same way. Accordingly, we choose no more than a single feature from each category. The final feature set is comprised of median log-pitch, HF500, and median vocal intensity. Spearman's rank-correlation coefficients between the selected features are computed per speaker in each database, and the median values are shown in Table 4; also, the median p-values for each database are reported. The correlations between the selected features are generally only medium-strength (0.4-0.7), which is comparable to the correlation strength between these feature's scores and the arousal labels. All of these correlations are significant at the $p < 10e-4$ level by the binomial proportion test.[2] If the correlations between features were stronger, diversity would be lower and we would expect only minor benefits from fusion. Further, we observe a case where one feature is corrupted by channel properties, but the others are robust to those variations. In emoDB, the vocal intensity is corrupted by varying mouth-to-mic distance, having a negative correlation with the other features; yet pitch and HF500 are still positively correlated with each other, and their produced scores are positively correlated with labeled arousal.

TABLE 5
Proposed Model Performance in Reference to Spearman's Rank Correlation, ρ_S, with Arousal Labels

	Corpus							
	IEMO-CAP		*emoDB*		*EMA*		*VAM*	
u-wt. fus. ρ_S	0.62		0.62		0.71		0.69	
wt. fus. ρ_S	**0.62**		**0.74**		**0.71**		**0.71**	
bin. UAR	**73%**		**84%**		**84%**		**77%**	
	ρ_S	*wt.*	ρ_S	*wt.*	ρ_S	*wt.*	ρ_S	*wt.*
log-pitch	0.50	0.78	0.75	0.83	0.54	0.79	0.56	0.76
HF500	0.45	0.81	0.69	0.67	0.65	0.87	0.37	0.63
vocal int.	0.59	0.90	−0.46	−0.09	0.67	0.87	0.65	0.80

Sig: All statistics significant at $p < 1e-4$.

3.2 Utterance-Level Ratings

In this section we present results of our system for rating vocal arousal per-utterance. It is assumed that each utterance has its own arousal label and that a set of neutral-labeled data is available; in Section 3.3 we will consider rating with no neutral-labeled data per speaker. In order to compare our approach (which produces a continuous rating) with a state-of-the-art approach (which produces a discrete rating), we must convert our rating to a binary decision. We select the intuitive threshold of 0 to convert our bounded rating into high and low arousal decisions. Unweighted average recalls (UAR) are reported for our model and the baseline approach. Results were also

[1] Loudness is computed with software described in Fernandez (2004) [44], which can be downloaded at <http://affect.media.mit.edu/software.php>.

[2] UAR is the average of the recalls for each class. When the classes are unbalanced, there is less confidence in the minority class. To account for this issue, we modify the binomial proportion test by setting the sample size, N, to be twice the size of the minority class. This produces a p-value which is more conservative.

TABLE 6
Comparison of the Proposed Model to State-of-the-Art Supervised Techniques with Unweighted Average Recall (UAR) in Predicting High/Low Arousal

	Corpus			
	IEMOCAP	*emoDB*	*EMA*	*VAM*
prop. model	**73%**	84%	84%	**77%**
baseline neutral-norm	62%	82%	86%	65%
baseline all-norm	72%	**90%**	**87%**	72%
baseline no-norm	59%	$50\%^{ns}$	71%	63%

Sig: ns- $p >= 0.05$; *else*- $p < 1e - 4$.

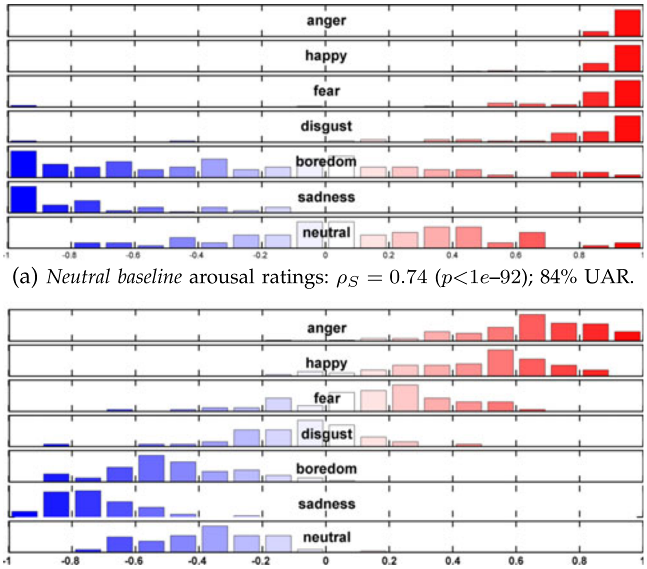
generated separately for the improvised and scripted portions of the IEMOCAP database, but only marginal differences were found.

Our arousal rating framework results in an accurate measure of vocal arousal on the four examined databases (Table 5). Spearman's rank-correlation coefficient between the proposed arousal rating and arousal labels is 0.62 for IEMOCAP, 0.74 for emoDB, and 0.71 for the other two databases when performing weighted fusion (with weights assigned as described in Section 2.3.3). Binary classification UARs with weighted fusion are 73-84 percent, well above chance (50 percent).

Each individual feature performs best in at least one database, producing a subrating that is most correlated with arousal labels. Also, fusion always produces a higher correlation than does any individual feature, except in the case of emoDB where one feature was corrupted.

Weighted fusion meets or exceeds performance of unweighted fusion (averaging) for all four databases. The increase with weighted fusion is most significant in the emoDB database. The intensity subrating has a medium negative correlation with arousal labels, because intensity is corrupted in the emoDB database by varying mouth-to-mic distance. Still, the weighted fusion framework assigns a very minor weight (-0.09) to intensity since it is not correlated with the average of the three features. This leads to an improvement from $\rho_S = 0.62$ (unweighted fusion) to $\rho_S = 0.74$.

The proposed vocal arousal rating compares very competitively with the state-of-the-art supervised approach (Table 6). Specifically, our proposed model is always better than the state-of-the-art no-normalization method ($p < 10e - 4$). Compared to the neutral-normalization method– which closely parallels our approach– our method is better in two cases (IEMOCAP and VAM; $p < 10e - 4$) and does not perform significantly differently in the other two ($p >= 0.05$). Furthermore, our model is competitive with the all-data normalization supervised approach; producing similar UAR in one case (IEMOCAP, $p >= 0.05$), higher UAR in one case (VAM, $p < 10e - 4$), and lower UAR in two cases (EMA, $p < 0.05$; emoDB, $p < 10e - 4$). For the emoDB database, the supervised approach outperforms our model by 6 percent UAR. This large difference may be dependent on assumptions of arousal-label distribution between databases; note that the neutral-normalization model produces similar performance (82 percent UAR, $p >= 0.05$) to our model and that the no-normalization model is at chance performance (50 percent). Our model outperforms the supervised approach for the VAM data ($p < 0.05$), the only natural emotional corpora we evaluate. The accurate vocal arousal decisions produced by our model indicate that these three features and our unsupervised fusion scheme can be as robust as a highly-tuned supervised classification approach that is prone to overfitting.



(a) *Neutral baseline* arousal ratings: $\rho_S = 0.74$ (p<1e–92); 84% UAR.

(b) *All-data baseline* arousal ratings: $\rho_S = 0.81$ (p<1e–100); 92% UAR.

Fig. 4. Histograms of emoDB arousal ratings for categorical emotions with differing speaker-baselines.

For further comparison, the UARs reported for this binary classification task are competitive with those presented by Schuller et al. (2010) [10] for emoDB; our result of 84 percent is approximately equal to their 75 percent quantile result of 82 percent UAR (Fig. 2b). In that study, different groupings of corpora were used for supervised training. Our model incorporates only three features compared to the 4,386 of the supervised approach. Thus, our model maintains interpretability while producing high UAR, and it has the further benefit of being a validated measure of scale-continuous arousal (Table 3).

The distribution of assigned arousal ratings for each emotional label in emoDB is displayed in Fig. 4 for both *neutral* and *all-data* (or global) speaker baseline models. The histograms clearly illustrate the differences between (i) neutral and (ii) all-data normalizations. (i) For neutral baseline modeling, feature values of each categorical emotion are compared to those of neutral data. In this framework, the Neutral data will receive feature-level arousal scores that are approximately uniformly distributed per-speaker, but in after fusion will tend to cluster around 0. For the emotions Anger, Happy, Fear, and Disgust (listed in decreasing order of arousal by visual inspection), there is a large concentration of arousal ratings near +1; this implies that the relevant feature values are often greater than the vast majority of observed Neutral features. Interestingly, Disgust was assigned a label of negative arousal (as in [10]), but it appears to evoke arousal that is higher than neutral, demonstrating the utility of this tool for unsupervised discovery. Boredom appears to be similar to Neutral, but slightly more negative. Finally, Sadness is the only emotion to receive primarily negative arousal ratings,

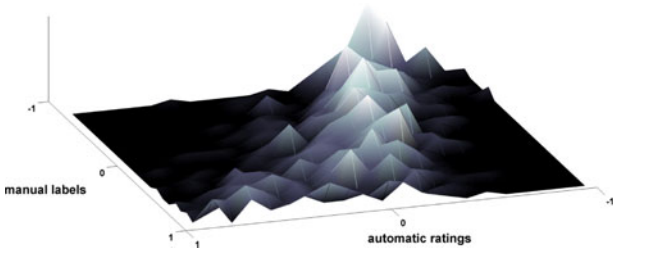


Fig. 5. Histogram of VAM arousal ratings. $\rho_S = 0.71$ ($p < 1e - 100$). Lighter color indicates higher occurence.

TABLE 7
Model Performance in Relation to Amount of Neutral Labeled Data Used for Speaker Baseline

	Corpus							
	IEMO-CAP		*emoDB*		*EMA*		*VAM*	
baseline data	ρ_S	*mad*	ρ_S	*mad*	ρ_S	*mad*	ρ_S	*mad*
all neutral	0.62	0.36	0.74	0.38	0.71	0.48	0.71	0.36
50% neutral	0.60	0.38	0.72	0.44	0.71	0.48	0.66	0.44
10% neutral	0.54	0.44	0.62	0.49	0.69	0.52	0.59	0.54
all data	0.60	0.32	0.81	0.51	0.71	0.52	0.65	0.28
neutral utterances per-speaker on average								
neutral count	111		8		74		4	
% of all utt.	16%		15%		23%		16%	

Performance in terms of Spearman's rank-correlation coefficient, ρ_S*, and mean-absolute difference (mad) with arousal labels. Note: All data equates to global normalization. Sig: All statistics significant at* $p < 1e - 4$.

indicating that the associated feature values often fall below the median Neutral features. The primary source of error (84 percent UAR) for a threshold of 0 would be from Disgust since it is primarily rated as high arousal.

(ii) For all-data baseline modeling, the same ordering of emotions by arousal rating is observed; in fact, the emotions seem to cluster by arousal rating. This improvement in ranking ($\rho_S = 0.81$ compared to $\rho_S = 0.74$) is likely due to having a wider variety of data, particular for the avoiding ceiling effects seen with neutral baseline modeling. We usually do not expect the all-data baseline model to classify binary arousal better than the neutral baseline, since neutral will shift its center depending on the distribution of emotions in the data. The observed shift towards low arousal brought the majority of Disgust instances below 0 arousal, which improved UAR to 92 percent. Again, our rating would suggest that Disgust in this data is actually a high arousal emotion; if it were labeled as such, the all-data model would perform much worse than the neutral model.

The relationship between arousal ratings and arousal labels in the VAM database is shown in Fig. 5; the prominent diagonal reveals the accuracy of the scale-continuous vocal arousal rating.

3.3 Effect of Baseline Data

It is important to test the performance of our system with varying amounts of neutral baseline data in order to understand how much is necessary to achieve a desired performance. It is reasonable to assume that a small amount of annotated neutral data will be available in many situations. Even so, the performance of the method without explicitly-labeled neutral data is investigated. The results of varying the amount and type of baseline data are displayed in Table 7. Performance metrics are Spearman's rank-correlation coefficient (ρ_S), a relative measure, and the mean-absolute difference (mad), an absolute measure. In each data set, arousal labels have been scaled to the range [-1, 1]. As a reminder, emoDB and EMA have discrete labels, so only scores of {-1, 0, 1} are possible; IEMOCAP and VAM have continuous labels.

With every database, more neutral data leads to increased correlations and decreased absolute error. However, significant correlations with labeled arousal are still achieved using as baseline only (on average): a single neutral instance in the emoDB and VAM databases; seven instances in the EMA database; and 11 instances in the IEMOCAP databases.

Surprisingly, for all-data normalization compared to all-neutral normalization, absolute error decreased in the two continuous-label arousal databases (VAM and IEMOCAP). This suggests that understanding the total range of data was more beneficial in assigning accurate arousal ratings than knowing where the center (neutral) was. Further, the relative performance using all-data baseline modeling is comparable to all-neutral baseline modeling across databases. This implies that the system can effectively rank the vocal arousal in a set of observed data from a single speaker, even without any labeling of that data; this is understandable since the three features which the system are built upon all correlate well with vocal arousal across databases and contexts. Overall, the model apparently needs very little labeled data to achieve impressive relative performance, yet high absolute accuracy in rating arousal is often not achieved.

3.4 Temporally-Continuous Arousal Rating

In this section the proposed vocal arousal rating model is extended to temporally-continuous arousal rating (one with a constant sampling rate). Our previous experiments only considered the case where the boundaries for an utterance were provided and a single label of arousal was given for each utterance. The primary benefit of this temporally-continuous rating is that it does not require any utterance boundaries. It provides a score for all voiced frames, and it assumes one speaker per recording.

Our approach to temporally-continuous arousal rating uses the same framework as in static ratings, but with minor adaptations. A decision is made every 10 ms using a sliding-window approach. Pitch, intensity, and HF500 are extracted with a 25 ms sliding-window; all unvoiced frames are assigned as a missing value (NaN in Matlab). Baseline models are created from all of a speaker's voiced frames within a single session (global baseline) to account for varied acoustic conditions between sessions. Since we are performing global normalization rather than neutral normalization, we may expect that the rating will correlate with the arousal labels, but that the absolute value of that rating will have less meaning. In short, we cannot say that a rating of 0.25 is necessarily positive arousal on average. Once the baseline model is available, scoring can be accomplished. Before scoring, each feature stream is smoothed with a 1-second

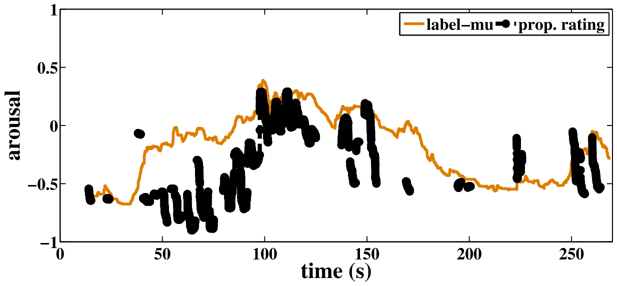


Fig. 6. Frame-based temporally-continuous arousal rating and arousal label of one CreativeIT recording.

(100-sample) median filter, ignoring missing values. Scoring and fusion are then completed as in the static scenario. As a final modification, the fused arousal rating is further smoothed with a 2-second (200-sample) median filter. Smoothing produced better, less-variable, ratings.

A result of this frame-based approach is shown in Fig. 6. Utterance-level decisions are performed in the same session and displayed in Fig. 7. (Utterance-level arousal labels were computed by taking the median of the arousal labels assigned to an utterance.) The frame-based ratings are highly variable, even within an utterance. The utterance-level ratings follow a similar course to the frame-based ratings, but may benefit from averaging information from multiple frames.

The performance of continuous-arousal rating is evaluated in terms of the Spearman's rank-correlation with continuous arousal tracking as well as mean-absolute-difference (mad). The arousal labeling was conducted by observing a video. A 2-second delay in annotation was assumed based on empirical evidence—this aligns with previously reported results on another annotated emotional database [46]. The frame-based approach produced correlations in the range $[-0.62, 0.89]$ as depicted in Fig. 8. It is clear that in some cases (especially near utterances with few voiced frames) the ratings have weak coupling to arousal labels. However, the ratings in the frame-based approach have a median correlation of 0.49, which is significantly greater than 0 by the Wilcoxon signed-rank test ($p < 1e-13$). Thus, the relative relationship between the vocal arousal rating and arousal labels is significant. However, as expected, the absolute values differ greatly. Specifically, the median session-level mean-absolute-difference between the automatic rating and the arousal labels is 0.44, which is much larger than the median session-level mean-subtracted-amplitude of the arousal labels (0.15). This implies that the error in predicting the arousal labels is larger than the variability of the labels themselves.

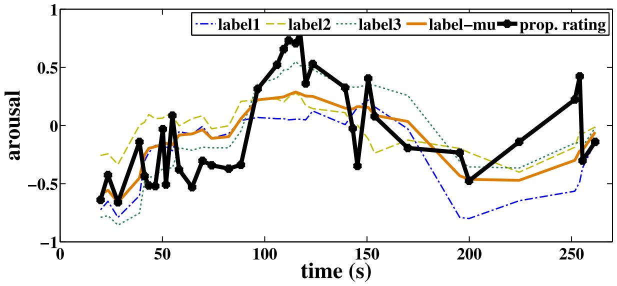


Fig. 7. Utterance-level (static) arousal rating and corresponding arousal label of one CreativeIT recording. *Note:* label-i is the label for the ith rater.

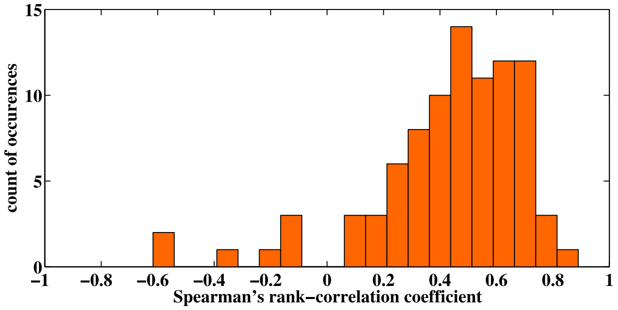


Fig. 8. Histogram of correlations between frame-based arousal ratings and arousal labels for the 90 CreativeIT speaker.

Static, utterance-level decisions are useful for comparison. The static method with global normalization produces median correlations in the range $[-0.27, 1]$ with a median of 0.50 (statistically significant at $p < 1e\text{-}15$). Thus, the two approaches produce comparable results in terms of relative arousal rating. The utterance-level approach is also a poor descriptor for absolute arousal, as expected from the global normalization. The median session-level mean-absolute-difference between rated and labeled arousal is 0.30, while the median session-level mean-subtracted-amplitude of labeled arousal is 0.22.

4 Discussion

An unsupervised (rule-based) approach to arousal detection is proposed and was tested across multiple corpora. The algorithm only requires baseline data, preferably from a neutral portion of speech. The algorithm assumes that changes in features relate to predictable changes in arousal.

4.1 Utterance-Level Vocal Arousal Ratings

The first step in re-designing our initial system proposed in Bone et al. (2012) [27] was to consider the addition of alternative features in Section 3.1. For instance, loudness was hypothesized as potentially more robust than intensity. While loudness did create higher correlations than intensity for emoDB, which has known energy issues, the two are approximately equal in performance on the other databases. Since loudness takes much longer to compute, and since our feature-score fusion framework essentially excluded intensity in emoDB, we selected intensity. We also chose other features suspected to contain orthogonal information: a measure of pitch and a measure of voice quality. Median pitch is clearly the highest performing pitch feature, as is HF500 for voice quality. Speaking rate was also investigated both from forced-alignment and from an intensity-based syllabic method. However, speaking rate does not produce consistent performance as hypothesized from relevant literature. We expect this is due to factors such as speaker idiosyncrasies in expression of emotion.

The algorithm, which automatically fuses the three chosen features, was applied to the various databases in Section 3.2. Medium-to-high Spearman's rank-correlation coefficients were observed for four emotional databases (IEMOCAP, emoDB, EMA, and VAM). Two important observations were made. First, each feature has the highest individual correlation with vocal arousal in at least one database; this suggests that a single feature is not optimal. Second, fusion improves correlation in all cases except emoDB; fortunately our fusion framework is successful in diminishing the effect of the corrupted vocal intensity feature. Thus, the utterance-level arousal ratings appear accurate and robust.

We further tested performance by comparing our proposed rating to a start-of-the-art supervised affect recognition method (openSMILE with linear SVM). The results support the use of this simple arousal rating framework for binary arousal classification. Our method achieves between 73 and 84 percent UAR. The only major differences between the supervised results and the proposed model exist for the emoDB and VAM databases. The supervised technique performed much better on emoDB, but emoDB is often regarded as a rather easy database for classification. In VAM, the only natural affective database that is examined, our method outperforms the supervised method. These results indicate that the proposed arousal rating is robust. Our method has the additional advantage of providing a verified scale-continuous measure, which is more precise than the binary measure by common state-of-the-art approaches. On another note, further support for incorporating baseline information is observed from the results of supervised classification with and without normalization.

4.2 Effect of Baseline Data

It is shown that the vocal arousal rating model does not always require large amounts of neutral data (Section 3.3). A generally observed trend is that more neutral data leads to higher correlations and lower mean-absolute errors, although the differences were rather small. Even with as little as 10 percent of the original neutral data (sometimes as low as 1 utterance per speaker) the arousal rating was still fairly successful. Furthermore, neutral data is not necessarily essential; normalization can be performed using all available data (referred to as global normalization). Global normalization leads to better performance in the emoDB database, but lower in the VAM. Overall, the results suggest very little labeled data is needed and potentially a small amount of unlabeled data will be sufficient—this will be discussed further in Section 4.4.

4.3 Temporally-Continuous Ratings

An extension of this method is proposed for rating the vocal arousal of data which does not have any utterance boundaries marked (Section 3.4). It is assumed that all data comes from the same speaker. Raw data from lapel microphones may fit this set of constraints.

The arousal rating using the frame-based approach with multiple smoothing criteria has medium correlation ($\rho_S = 0.49$) with manual annotations in the CreativeIT database. For comparison, the utterance-boundaries can also be used to produce the utterance-level ratings as before; this produced a correlation of ($\rho_S = 0.50$), which is very similar to the frame-based result (although the mean-absolute error was lower for the utterance-level approach). Thus, continuous vocal arousal rating is achievable in multiple ways using the proposed framework.

4.4 Guidelines for Application

The arousal rating framework described in this article is intended for interdisciplinary use. Researchers can obtain the vocal arousal rating software in Matlab format from http://sail.usc.edu/sail_tools.php.

Several guidelines for use are suggested:

Speech data. First, each speaker's data is assumed to have come from only a single speaker and from identical acoustic settings. Praat uses a Viterbi decoding scheme for pitch extraction which assumes that only a single speaker's voice is processed. Second, no changes in environment or recording parameters should occur. If incorporating data from the same speaker over multiple sessions, it is suggested that each session be treated separately– in the software this can be accomplished by creating separate speaker IDs.

Baseline data. Each speaker requires some amount of baseline data. The data should be from a neutral-labeled portion of speech if possible. This is because global normalization is less well-motivated and needs further experimentation to be better understood.

Interpretation. This measure of vocal arousal should most often be treated as a relative measure, not an absolute one. A potential application would be to model the dynamics of vocal arousal across a session jointly with another temporal sequence (i.e., autonomic arousal or a tagged event). While this measure produced state-of-the-art performance in binary arousal detection across corpora, it is important to point out that this method will produce different absolute numbers for vocal arousal depending on the baseline data. For example, high arousal data used as baseline will lead to arousal ratings that must be interpreted differently than with neutral data as baseline. However, shifts in the arousal baseline are not expected to have a great effect on the relative relevance of the measure (such as with correlation).

5 Conclusion

A simple measure of vocal arousal utilizing only three features was proposed and validated on multiple affective databases. The measure is designed to fill a void between demands of behavioral studies and the lack of simple, robust measures available from the engineering community. The framework assumes that some neutral baseline data is available; moderate robustness to this assumption is demonstrated in the performance of global speaker-normalization.

The proposed framework can be applied to other tasks as well, assuming there are robust correlates of the target dimension. These descriptors are currently lacking for valence [47]. Empirical evidence suggests valence recognition is a more difficult task (e.g., [9], [10]), and some researchers think valence is dependent on voice quality [35], context, or other communicative modalities.

In the future, this arousal rating will be applied to other Behavioral Signal Processing (BSP) scenarios as a dynamic measure of arousal, including in domains such as couple therapy [48] and autism [49], [50], [51]. In particular, it will be used to model the temporal evolution of affect between interlocutors.

ACKNOWLEDGMENTS

The authors would like to thank the researchers who made their databases available for our study. This work was supported by funds from NSF and NIH. The first author was also supported by the Achievement Rewards for College Scientists Foundation.

REFERENCES

[1] W. Schramm, "How communication works," in Schramm, W. (Ed.), The Process and Effects of Communication, University of Illinois Press, Urbana, IL, *Mass Media Soc.*, p. 51, 1954.

[2] M. L. Patterson, "An arousal model of interpersonal intimacy," *Psychol. Rev.*, vol. 83, no. 3, p. 235, 1976.

[3] E. Duffy, "The psychological significance of the concept of arousal or activation," *Psychol. Rev.*, vol. 64, no. 5, p. 265, 1957.

[4] P. L. Harris, *Children and Emotion: The Development of Psychological Understanding*. Oxford, U.K.: Blackwell, 1989.

[5] M. Hernandez-Reif, T. Field, M. Diego, and M. Ruddock, "Greater arousal and less attentiveness to face/voice stimuli by neonates of depressed mothers on the brazelton neonatal behavioral assessment scale," *Infant Behav. Develop.*, vol. 29, no. 4, pp. 594–598, 2006.

[6] R. P. Hobson, J. Ouston, and A. Lee, "Emotion recognition in autism: Coordinating faces and voices," *Psychol. Med.*, vol. 18, no. 4, pp. 911–923, 1988.

[7] G. Dawson, S. J. Webb, L. Carver, H. Panagiotides, and J. McPartland, "Young children with autism show atypical brain responses to fearful versus neutral facial expressions of emotion," *Develop. Sci.*, vol. 7, no. 3, pp. 340–359, 2004.

[8] J.-A. Bachorowski, "Vocal expression and perception of emotion," *Current Directions Psychol. Sci.*, vol. 8, no. 2, pp. 53–57, 1999.

[9] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?" in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 1553–1556.

[10] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul.–Dec. 2010.

[11] B. Z. Pollermann, "A place for prosody in a unified model of cognition and emotion," in *Proc. 1st Int. Conf. Speech Prosody*, 2002, pp. 17–22.

[12] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The Panas scales," *J. Personality Soc. Psychol.*, vol. 54, no. 6, p. 1063, 1988.

[13] C. S. Carver and T. L. White, "Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales," *J. Personality Soc. Psychol.*, vol. 67, no. 2, p. 319, 1994.

[14] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008.

[15] 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 *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interaction Workshops*, Sep. 2009, pp. 1–7.

[16] M. Nowak, J. Kim, N. W. Kim, and C. Nass, "Social visualization and negotiation: Effects of feedback configuration and status," in *Proc. ACM Conf. Comput. Supported Cooperative Work*, 2012, pp. 1081–1090.

[17] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Proc. IEEE Workshop Autom. Speech Recog. Understanding*, 2001, pp. 240–243.

[18] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2003, vol. 2, pp. II-1–II-4.

[19] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interfaces*, 2004, pp. 205–211.

[20] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–302, Mar. 2005.

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

[22] 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 Commun.*, vol. 53, pp. 1062–1087, 2011.

[23] H. J. Steeneken and J. H. Hansen, "Speech under stress conditions: Overview of the effect on speech production and on system performance," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 1999, vol. 4, pp. 2079–2082.

[24] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-corpus classification of realistic emotions-some pilot experiments," in *Proc. 3rd Inter. Workshop Emotion*, 2010, pp. 77–82.

[25] P. Juslin and K. Scherer, "Vocal expression of affect," in *The New Handbook of Methods in Nonverbal Behavior Research*. Oxford, U.K.: Oxford Univ. Press., 2005, ch. 3, pp. 65–135.

[26] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "On the influence of phonetic content variation for acoustic emotion recognition," in *Proc. Perception Multimodal Dialogue Syst.: 4th IEEE Tutorial Res. Workshop Perception Interactive Technol. Speech-Based Syst.*, 2008, pp. 217–220.

[27] D. Bone, C.-C. Lee, and S. S. Narayanan, "A robust unsupervised arousal rating framework using prosody with cross-corpora evaluation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1175–1178.

[28] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self-assessment manikins," in *Proc. IEEE Workshop Autom. Speech Recog. Understanding*, 2005, pp. 381–385.

[29] 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. Lang. Resources Eval.*, vol. 42, pp. 335–359, 2008.

[30] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 497–500.

[31] J. Kim, S. Lee, and S. S. Narayanan, "An exploratory study of the relations between perceived emotion strength and articulatory kinematics," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 2961–2964.

[32] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2005, pp. 1517–1520.

[33] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, S. Narayanan, and D. Tx, "The USC creativeit database: A multimodal database of theatrical improvisation," in *Proc. Multimodal Corpora Workshop: Adv. Capturing, Coding Anal., Multimodality*, 2010, pp. 64–68.

[34] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2008, pp. 865–868.

[35] K. Scherer, "Vocal affect expression: A review and a model for future research," *Psychol. Bull.*, vol. 99, no. 2, p. 143, 1986.

[36] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behav. Res. Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[37] M. Young, M. Landy, and L. Maloney, "A perturbation analysis of depth perception from combinations of texture and motion cues," *Vis. Res.*, vol. 33, pp. 2685–2696, 1993.

[38] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[39] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 312–315.

[40] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 3201–3204.

[41] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, 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 *Proc. Interspeech*, 2013, pp. 148–152.

[42] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[43] Z. Zhang, F. Weninger, M. Wollmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011, pp. 523–528.

[44] R. Fernandez, "A computational model for the automatic recognition of affect in speech," Ph.D. dissertation, Dept. Archit., Massachusetts Institute of Technology, Cambridge, MA, USA, 2004.

[45] E. Zwicker, H. Fastl, U. Widnmann, K. Kurakata, S. Kuwano, and S. Namba, "Program for calculating loudness according to din 45631 (iso 532b)," *J. Acoust. Soc. Japan (E)*, vol. 12, no. 1, pp. 39–42, 1991.

[46] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 85–90.

[47] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1179–1182.

[48] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features," *Speech Commun.*, vol. 55, no. 1, pp. 1–21, 2013.

[49] 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 *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 1497–1500.

[50] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *J. Speech, Lang., Hearing Res.*, (in press) 2014.

[51] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1043–1046.

Daniel Bone (S'09) received the BS degrees with highest distinction and honors in electrical engineering and computer engineering from the University of Missouri-Columbia, in 2009, and the MS degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2011. He is currently pursuing the PhD degree in the Signal Analysis and Interpretation Laboratory (SAIL) at USC. His research interests concern human-centered signal processing and machine learning, with a focus on creating engineering techniques and systems for societal applications in human health and well-being. Specifically, he aims to develop computational methods for characterizing and eventually informing intervention of neurodevelopmental disorders such as Autism (population prevalence of 1 in 88). He is a member of the IEEE Signal Processing Society. He received the Alfred E. Mann Innovation in Engineering Fellowship 2014, the Achievement Rewards for College Scientists Scholarship 2012-2014, the NSF GK-12 Fellowship 2012-2013, and the USC Annenberg Fellowship 2009-2011. He led a team that won the Interspeech 2011—Intoxication Sub-Challenge. He is a member of Eta Kappa Nu. He is a senior member of the IEEE.

Chi-Chun Lee (M'13) received the BS degree with honor, magna cum laude, in electrical engineering from the University of Southern California (USC), Los Angeles in 2007, and the PhD degree in electrical engineering from the University of Southern California in 2012. He is an assistant professor at the Electrical Engineering Department of the National Tsing Hua University (NTHU), Taiwan. His research interests are in human-centered behavioral signal processing, emphasizing the development of computational frameworks in recognizing and quantifying human behavioral attributes and interpersonal interaction dynamics using machine learning and signal processing techniques. He has been involved in multiple interdisciplinary research projects and has conducted collaborative research with researchers across domains of behavior science. He was awarded with the USC Annenberg Fellowship (2007-2009). He led a team to participate and win the Emotion Challenge-Classifier Sub-Challenge in Interspeech 2009. He is a coauthor on the best paper award in Interspeech 2010. He is a member of Tau Beta Pi, Phi Kappa Phi and Eta Kappa Nu. He is a member of the IEEE.

Shrikanth Narayanan (M'95-SM'02-F'09) is the Andrew J. Viterbi professor of engineering at the University of Southern California (USC), Los Angeles, and holds appointments as professor of electrical engineering, computer science, linguistics, and psychology and as the founding director of the Ming Hsieh Institute. Prior to USC, he was with AT&T Bell Labs and AT&T Research from 1995 to 2000. At USC, he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered information processing and communication technologies with a special emphasis on behavioral signal processing and informatics. He is a fellow of the Acoustical Society of America and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also an editor for the *Computer Speech and Language Journal* and an associate editor for the *IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, and the *Journal of the Acoustical Society of America*. He was also previously an associate editor of the *IEEE TRANSACTIONS OF SPEECH AND AUDIO PROCESSING* (2000-2004) and the *IEEE SIGNAL PROCESSING MAGAZINE* (2005-2008). He received a number of honors including Best Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and in 2009 (with C. M. Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010-2011. He has published more than 500 papers and has 15 granted U.S. patents. He is a fellow of the IEEE.