An Analysis of the Relationship between Signal-derived Vocal Arousal Score and Human Emotion Production and Perception

Chi-Chun Lee[1], *Daniel Bone*[2], *Shrikanth S. Narayanan*[2]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA

cclee@ee.nthu.edu.tw, dbone@.usc.edu, shri@sipi.usc.edu

Abstract

Bone *et al.* recently proposed an unsupervised signal-derived vocal arousal score (VC-AS) based on fusion of three intuitive acoustic features, i.e., pitch, intensity, and HF500, and have shown the effectiveness of quantifying human perceptual ratings of arousal robustly across multiple corpora. Due to the readily-applicable nature of the system, this objective quantification scheme could foresee-ably be used in multiple fields of behavioral science as an objective measure of affect. In this work, we investigate in detail the relationship of this signal-derived measure to both intended arousal expression (i.e., production aspect) and perceived arousal rating (i.e., perception aspect). On the perception side, our results on three databases (EMA, VAM, and IEMOCAP) indicate that VC-AS agrees with mean perception at least as well as an average individual rater does. Regarding production, we observe that intended arousal correlates more with VC-AS than mean perception (EMA and IEMOCAP), and that VC-AS correlates more with intended arousal than perceived arousal (EMA); these findings are surprising given that the framework is motivated by extensive affective perception studies, although there is physiological backing. Implications for the use of VC-AS for novel scientific study (e.g., to mitigate subjectivity) is further discussed.

Index Terms: vocal arousal rating, affective perception, affective production

1. Introduction

Emotion is one of the most fundamental attributes in governing human behaviors. In the past decade, the field of affective computing has produced a collection of works investigating various aspects of objectively modeling human emotion [1]. There is particularly expansive literature on recognizing human emotion from measurable signals, e.g., speech, gesture, and text [2]. These works have led to significant advancements in the design of human-machine interfaces: e.g., virtual humans [3], natural dialog interfaces [4], and intelligent tutoring systems [5]. Recently, the interdisciplinary field of Behavioral Signal Processing, BSP [6], has emerged. BSP models high-level human behavioral constructs using novel computational frameworks in order to support and supplement a domain-expert's decision-making process. Applications in mental health include: autism spectrum disorder [7], addiction [8], and couple's therapy [9].

Modern emotion recognition systems are capable of achieving high recognition rates within a database. However, a key component on which many domain experts rely, but remains lacking, is robust application across databases and scenarios while maintaining interpretability. Instead of the conventional supervised learning techniques, which are prone to data-overfitting, Bone *et al.* presented a vocal arousal scoring method (VC-AS) that incorporates a minimal set of knowledge-inspired vocal features into an unsupervised (rule-based) system design [10]. The framework is interpretable, scale-continuous (as opposed to usual discrete *n*-class arousal states), and operational without much (if any) manual human labeling. While simple, the framework has demonstrated high correlation to mean human ratings of emotional arousal (a.k.a., activation).

VC-AS is a BSP measure of emotional arousal (e.g., [11]). It not only could help advance robust human-machine interface design, but also create opportunities to quantitatively investigate human emotion production and perception mechanisms from a completely objective perspective. Human interaction can be conceptualized as a communication system, i.e., a production–perception pair. On the production side, humans encode affective information in behaviors; and on the perception side, humans decode that affective information in order to respond appropriately. Certain past works have investigated the scientific underpinning of emotional behavior production [12, 13]; the quantity of work regarding of mechanisms of human affective perception is much larger (e.g., meta studies on emotional perception [14, 15, 16]).

VC-AS, as a signal-derived vocal arousal measure, naturally fits in the empirical production-perception study paradigm (mainly due to the fact that the framework does not depend on human perceptual labeling in order to *train* the system). It has benefits compared to using human judgments (i.e., self report or observer report) as measures of affect, particularly for emotional arousal. Due to the high correlation to mean perceptual ratings done by humans, now we can imagine VC-AS as a *valid rater* by itself, an *evaluator* derived objectively from signal. We can investigate the following questions quantitatively:

1. How does the agreement between VC-AS and raters compare to how well the raters agree among themselves?
2. Which part of this human emotional communication system does the VC-AS signal capture more, production or perception? And how well does target production agree with VC-AS versus human perception?

We investigate Question 1 in three emotional databases and find that VC-AS has similar performance to an average rater: in one case falling slightly below inter-rater agreement, in another far exceeding it. We answer Question 2 from two databases, finding that VC-AS is a better measure of target production than human perception– possibly indicating that VC-AS more closely relates to intended arousal production than perception. Our initial findings will support future experiments using VC-AS as a scientific-discovery tool.

The following paper is organized as follow: section 2 de-

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

Corpus	**Style**	**Emotion**	**Intended Emo?**	**Label**	$-$	$+$	Neu	**Total**	**Speakers**	**Setting**	**Language**
IEMOCAP	improv.	acted	improv. role	ordinal	*N/A*	*N/A*	(1112)	6883	10 (5f,5m)	studio	English
EMA	read	acted	script	categorical	408	338	221	967	3 (1f,2m)	studio	English
VAM	spont.	natural	*N/A*	continuous	*N/A*	*N/A*	*N/A*	947	47 (32f,15m)	noisy	German

scribes our research methodology; section 3 provides experimental results and discussions; finally section 4 concludes with future directions.

2. Method

In this section we describe the studied affective databases, vocal arousal score (VC-AS), and computational analysis techniques.

2.1. Affective Databases

Our experiments are conducted on three publicly available affective speech databases: IEMOCAP, EMA, and VAM (Table 1). Databases were selected primarily based on availability of perceptual rating data; secondarily, we sought some measure of target production. These databases are diverse: scripted and spontaneous, English and German, acted and natural. The databases were used in our previous work with the vocal arousal score (VC-AS) [10, 17].

IEMOCAP contains mixed-gender spontaneous improvisation as well as scripted interactions between trained actors [18] with both continuous and categorical arousal rating. There are 5 dyads (10 speakers), and 6905 utterances in total: 2388 utterances from improvisation and 4517 from scripted interaction; these numbers are post-exclusion for overlapped speech, very poor audio quality, and data with no voiced frames. Arousal perception was performed by six raters on a 5-point integer scale, which is then averaged to obtain a final rating; in practice, 2-3 raters scored each utterance. Raters tagged data sequentially using audio and visual information.

USC-EMA is comprised of English read affective speech from three trained actors performing five categorical emotions–neutral, hot anger, cold anger, happiness, and sadness. Actors were requested to read lexically-neutral utterances after immersing themselves in the required affective state; additionally, the actors spoke in three speaking styles: normal, loud, and fast. EMA stands for electro-magnetic articulography; the actors additionally had sensors on their face and tongue for speech production research [12, 19]. Data is perceptually evaluated by 4-5 raters. For purposes of comparison with VC-AS, we map hot anger and happiness to high arousal and cold anger and sadness to low arousal; this is supported by Figure 1 in [17].

VAM is a natural emotion speech database of dyadic and triadic interactions recorded during a German TV talk-show "Vera am Mittag" (Vera at noon). In total, there are 47 distinct speakers in the audio release and a total of 947 utterances. Since we must have enough data to develop a baseline for a speaker, we concentrate on only 37 speakers who spoke at least 10 utterances (μ=24 utterances). Each utterance is scored on a continuous-scale in the range 0 to 1 for valence, activation (arousal), and dominance by 7 to 17 raters.

Production, or target, emotion information is difficult, if not impossible, to find for natural emotional corpora. One would have to infer the appropriate behavioral response for a person in a certain situation, or rely on self- or observer-perceptual report. We attempt to obtain measures of production for the two acted databases. *EMA* is rather straightforward to obtain a target production for, since each utterance is associated with a prescribed emotional goal. We note that there can be errors in production, but at least the actors should have been motivated to produce this emotional expression. *IEMOCAP* has a pseudo-target of production which we present for tentative analysis. The improvisation portion of the database prescribes a general emotional state to a speaker based on their role in the scenario [18]. We assign to each utterance for that speaker, the arousal mapping of their categorical affective role. For example, if a speaker's role is to be 'frustrated', we map the target production for all utterances to 'low arousal'. Since we could not find meta-data indicating which speaker assumed which of two roles, we used a heuristic where we identified the speaker's role based on matching perceived arousal to target arousal. The assignment is only made if 50% or more of utterances match, otherwise we reject utterances; after rejection, we are left with 1318 utterances for analysis.

2.2. Vocal Arousal Score (VC-AS) Computation

We have designed VC-AS as an unsupervised system for vocal arousal rating from the speech signal (shown in Figure 1) that provides state-of-the-art performance with perceived affect [10, 17][1]; it is unsupervised in the sense that no affective labeling is required since the system relies on rules. Certain acoustic cues have been shown very reliable across many experiments; using this information, we design a system which takes as input feature values for an individual speaker, and gives as output an arousal score associated with each utterance. In essence, the system rank-orders three reliable acoustic correlates of arousal, then fuses the ordered-values for robustness.

Our acoustic features were selected based on the summative analysis of a multitude of perceptual and engineering recognition experiments [14] and our own empirical validation [10]. Specifically, our system utilizes median pitch (f_0), median intensity, and median HF500, all of which have a positive relation with arousal; for example, a person's f_0 increases during periods of excitement (increased arousal). Pitch and intensity are extracted using Praat [20]. HF500 is a voice quality ratio of the spectral energy above 500 Hz to that below.

Based on these primary assumptions, or rules, we create an unsupervised system. However, we first need to know what a

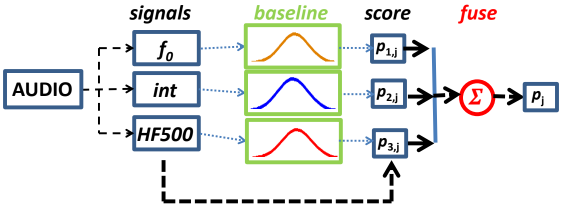

Figure 1: *Arousal rating system diagram showing progression from raw data (utterance 'j'), to features, to individual feature scores, and finally to fused score p_j.*

[1] VC-AS can be downloaded at `http://sail.usc.edu/~dbone`

speaker's baseline is, for without the baseline, a feature value has no absolute meaning; for example, Is 150 Hz pitch indicative of high arousal? We can model all of a person's speech as baseline, although this leads to arousal scores in which the relative ordering is meaningful, but the absolute values are not. Instead, if we have a small sample of "neutral" data to use as a baseline, we enhanced interpretation of VC-AS raw values.

The scoring of our system is as follows. First, we obtain baseline models for each speaker, which are simply a collection of feature values. Then, for each feature $x_{i,j}$ of utterance j to be scored, individual feature score $p_{i,j}$ (in the range [-1,1]) is calculated against a baseline 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. Ratings are then fused by normalized weighted summation, with weights based on correlation of individual feature scores with the unweighted mean [10].

2.3. Score and Rater Agreement Analysis

In this work, agreement between raters is compared to agreement with VC-AS in order to assess the quality of *VC-AS* as a replacement or additional *rater*. We compute two intra-rater agreement measures, quantified through Spearman's correlation: (i) agreement between individual raters and the mean of the remaining raters, and (ii) mean agreement between individual raters and other individual raters. Averages and standard deviations across raters are reported. In each experiment, VC-AS agreement is computed just as if it were another rater. We also measure the relationship between perceptual measures and measures of production using Spearman's rank-correlation.

VC-AS is most reliable for within-speaker analyses due to sensitivity to baseline validity for intra-speaker analyses. Specifically, using all samples as baseline (global normalization) assumes the true arousal distributions are similar across speakers; and neutral sample normalization necessitates a representative sample size. Since neutral normalization is most generalizable inter-speaker, we utilize it for all but EMA data. VAM data do not contain any categorical labels and there are relatively few utterances per speaker; we ground our intra-speaker analysis by selecting utterances with mean perceptual ratings near '0' as described in [17]. For IEMOCAP, the affective distribution from a speaker may vary greatly, so we also use neutral normalization from categorical emotional labels. For EMA, we do use global normalization because we have many samples per speaker and they all contain the same emotional distribution.

3. Results and Discussion

3.1. VC-AS and Human Perception of Affect

Essentially, by imagining VC-AS as a rater itself, we first attempt to understand how well VC-AS, which is a machine-based rater, evaluates a *subjective* attribute, i.e., emotional arousal, compared to human-based raters evaluating the same *subjective* attribute. In this experiment, we investigate the following question (Question 1 mentioned in the introduction).

1. How does the agreement between VC-AS and raters compare to how well the raters agree among themselves?

Table 2 summarizes our correlation results (statistical testing is based on Fisher r-to-z transform). From our previous work, we have determined that VC-AS approximates the perceptual mean ratings of emotional arousal well– in fact, as well as state-of-the-art supervised methods do in cross-corpora analysis. In answering Question 1, we further demonstrate that VC-AS, as a machine-based rater, can achieve comparable agreement with humans as compared to humans raters themselves. In fact, for databases with a limited number of raters (i.e., IEMOCAP), our analysis even shows that human agree with VC-AS more than human agrees among themselves on average. This insight has strong implication for behavioral experts performing analysis tasks involving perceiving and observing the emotional arousal for the subject of interest. In this case, VC-AS is a viable alternative approach (compared to observational coding) in mitigating the issue around subjectivity while still obtaining a robust and accurate measure, albeit it narrow definition of arousal.

There is one database for which humans agree among themselves statistically more than they agree with VC-AS: VAM (note that this trend exists for EMA, but it is not statistically-significant). First, we note that there are many more raters for VAM than other data; therefore, in terms of carrying out statistical testing, significance can be achieved in this test with smaller effect sizes. We can further explain the differences by the small number of utterances per-speaker in the data (min=10, μ=24). It is difficult for our unsupervised system to be successful in intra-speaker analysis given so few samples; in particular, we have difficulty finding *either* a speaker's neutral (neutral-baseline) *or* observing the full range of a speaker's arousal expression (global-baseline). Since we compute statistics across all data (i.e., across speakers) but compute VC-AS within-speaker (i.e., speaker dependent), the raw scores of our VC-AS may not generalize across speakers. In fact, if we conduct the same correlation experiment in a within-speaker setting and compute the correlation between the *number of utterances* from a particular speaker and the *correlation of VC-AS* with the *mean human perceptual ratings*, there exists a statistically significant positive relationship (ρ_S=0.35, p<0.04). Presumably there are also more neutral (baseline) utterances, which supports that our system works better when we have more baseline data.

Another interesting observation we see is that with the availability of more human ratings, the individual idiosyncrasy (subjectivity in emotion labeling) seems to be averaged out; the emotional perceptual rating is hence *cleaner* and more *reliable* (evident in VAM). However, if there are only a few raters, VC-

Table 2: *Agreement between signal-derived VC-AS and perceptual ratings in terms of Spearman's rank-correlation (ρ_S). Values are presented as mean (stdv.) across raters. The statistically larger value (α=0.05) is* ***bolded***. *Legend: Ind. Rater = individual rater; Mean Rating = speaker-independent mean of individual ratings.*

		EMA		VAM		IEMOCAP	
Measure	**Reference**	*Agreement*	*Stat. Diff.*	*Agreement*	*Stat. Diff.*	*Agreement*	*Stat. Diff.*
Ind. Rater	*Mean Rating*	0.61 (0.031)	p=0.39	**0.74** (0.055)	p<0.05	0.51 (0.038)	p<0.05
VC-AS	*Mean Rating*	0.60 (0.018)		0.70 (0.005)		**0.67** (0.034)	
Ind. Rater	*Ind. Rater*	0.49 (0.036)	p=0.45	**0.62** (0.051)	p<0.05	0.43 (0.023)	p<0.05
VC-AS	*Ind. Rater*	0.47 (0.058)		0.58 (0.063)		**0.57** (0.050)	
Number of Raters		5		17		3	

AS can provides a *reliable* and *consistent* arousal estimate, i.e., both a high correlation to mean perceptual arousal rating (reliable) and a higher agreement with humans as compared to humans among themselves (consistent). For example, EMA has 5 raters and IEMOCAP has 3 raters, and in both cases, an individual rater agrees with VC-AS either more significantly than (IEMOCAP) or without any statistical difference to (EMA) raters among themselves. This result is quite promising since in most real life applications, especially mental health-related, access to experts raters of internal arousal state is often limited. VC-AS can help mitigate much of the concern in controlling for subjectivity in the process of emotion labeling without recruiting a substantial number of raters to perform observational coding (a common practice in behavioral science studies).

For further insights, we analyze within-speaker agreement in EMA database (Table 3), where it is clear that VC-AS performs at the same level of ambiguity as human perceptual raters. Speaker 1's emotional expression may be more *prototypical* as compared to Speaker 2 and Speaker 3. This is reflected not only in the agreement numbers of subjective human ratings but also in the numbers associated with VC-AS. To further understand the VC-AS utility, we analyze intended arousal.

Table 3: *Agreement between signal-derived VC-AS and perceptual ratings in terms of Spearman's rank-correlation (ρ_S) in EMA (within-speaker analyses). Statistically larger values (α=0.05) are* ***bolded****. Legend: Indv. Rater = individual rater; Mean Rating = speaker-independent mean of indv. ratings.*

		Speaker 1	**Speaker 2**	**Speaker 3**
Measure	**Reference**	*Agreement*	*Agreement*	*Agreement*
Indv. Rater	*Mean Rating*	0.697	0.642	0.535
VC-AS	*Mean Rating*	0.760	0.579	0.469
Indv. Rater	*Indv. Rater*	0.630	0.523	0.410
VC-AS	*Indv. Rater*	0.669	0.466	0.361
Indv. Rater	*Target Arousal*	0.683	0.601	0.437
VC-AS	*Target Arousal*	**0.835**	0.630	**0.572**

3.2. VC-AS and Target Production of Affect

On the side of production, since VC-AS is computed from the *behavioral signals* captured from the subject's production, we can further analyze whether this *signal* captures more of the subject's *intended emotion* when compared to the human rater's perceptual decoding of the subject's *intended emotion*. Our second question of interest is:

2. Which part of this human emotional communication system does the VC-AS signal capture more, production or perception? And how well does target production agree with VC-AS versus human perception?

Results for EMA are shown in Table 3. We observe strong evidence that, on average, the correlation between VC-AS and a speakers' intended arousal expression (production) is higher than the correlation between individual raters' perceptual evaluation and the intended arousal. This result is quite unique because the framework of VC-AS is largely developed based on many perceptual studies and emotion recognition works (which are also mostly done by training systems to recognize human-based perceptual emotion labels). While some of the features we derive carry physiological motivation, it is interesting to note that given the conceptualization of a production-perception pair and imagining VC-AS as a rater itself, this machine-based rater is closer to the side of production than human raters are - indicating that it might capture more of the subject's *intent* as compared to the human rater's perceptual decoding tendencies.

For IEMOCAP, we compute the correlation between an actor's target arousal and both perceived arousal and VC-AS; we find that the correlation between production and VC-AS is larger than between production and mean perception: ρ_S=0.39 vs. ρ_S=0.25, respectively (p<1e-4). The result seems to point to the same finding as in the EMA database, though a caveat should be made that the labels of *intended emotion* are not as explicit as we have within EMA data.

In answering Question 2, we further demonstrate that, treating VC-AS either as a signal-based measure of arousal or as a valid machine-based arousal rater, VC-AS seems to be capable of capturing more of the emotional production information than an individual human arousal rater is able to (on average). This not only further strengthens the argument that VC-AS provides a valid quantitative approximation of emotional arousal, but also details the informational content that the VC-AS signal holds in the emotion production-perception paradigm of human communication (VC-AS receives less distortion from the communication channel of arousal transmission versus the human receiver, or rater, does). This result, while preliminary, could open up new opportunities for behavior scientists to investigate scientific questions related to emotional behavior production by leveraging VC-AS as a direct quantitative measure.

4. Conclusion

As interdisciplinary efforts in understanding and modeling human behavior increase, a readily-applicable, interpretable, and robust emotional arousal measure could provide domain experts a new computational framework in their unique studies. Bone *et al.* recently proposed a vocal-based arousal measure (VC-AS) that achieves reliable results across multiple databases. In this work, we conduct additional experiments under the emotional production-perception paradigm of human communication. In our first analysis related to the side of perception, we demonstrate that VC-AS (treated as a rater) can achieve comparable agreement level with the mean of human raters as compared to the average agreement computed among human raters. Furthermore, if there are only a small number of human raters available, VC-AS provides an advantageous, i.e., more consistent, approach to perceptual quantification of emotional arousal. In our second analysis, relating to the aspect of production, we show that VC-AS captures more information about the target emotion than a human rater does, although VC-AS is largely motivated by perceptual studies. The reason could possibly be due to the fact that VC-AS is computing critical signals directly from the speakers, capturing cues modulated primarily by intent. These findings not only reinforce the robustness of VC-AS, but also position the use of VC-AS in human emotion interaction study to better tease apart the production-perception pair in an objective, i.e., signal-based, manner.

There are multiple directions of future research. One of the immediate goals is to gather additional databases where the intended emotion is available in order to further substantiate our findings in Section 3.2. Further, VC-AS, aside from its obvious usage in robust emotion recognition, can provide an objective measure of human internal arousal states. Moreover, we will leverage VC-AS as a quantitative measure in the study of emotional interplay during human social interactions by working closely with the appropriate domain experts.

5. Acknowledgment

The research is supported in part by funds from the Ministry of Science and Technology and the National Science Foundation.

6. References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[3] J. Gratch and S. Marsella, "Tears and fears: Modeling emotions and emotional behaviors in synthetic agents," in *Proceedings of the fifth international conference on Autonomous agents*. ACM, 2001, pp. 278–285.

[4] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, 2005.

[5] K. D. Sidney, S. D. Craig, B. Gholson, S. Franklin, R. Picard, and A. C. Graesser, "Integrating affect sensors in an intelligent tutoring system," in *Affective Interactions: The Computer in the Affective Loop Workshop at*, Conference Proceedings, pp. 7–13.

[6] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. PP, no. 99, pp. 1–31, 2013.

[7] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1162–1177, 2014.

[8] D. Can, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features." in *INTERSPEECH*, 2012.

[9] 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 Communication*, vol. 55, no. 1, pp. 1–21, 2013.

[10] D. Bone, C.-C. Lee, and S. Narayanan, "Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 201–213, 2014.

[11] D. Bone, C.-C. Lee, A. Potamianos, and S. Narayanan, "An investigation of vocal arousal dynamics in child-psychologist interactions using synchrony measures and a conversation-based model," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] 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.

[13] J. Kim, S. Lee, and S. Narayanan, "A study of interplay between articulatory movement and prosodic characteristics in emotional speech production," in *Proceedings of Interspeech*, Conference Proceedings, pp. 1173–1176.

[14] P. N. Juslin and K. R. Scherer, "Vocal expression of affect," *The new handbook of methods in nonverbal behavior research*, pp. 65–135, 2005.

[15] J. A. Russell, "Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies." *Psychological bulletin*, vol. 115, no. 1, p. 102, 1994.

[16] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *Affective Computing, IEEE Transactions on*, vol. 4, no. 1, pp. 15–33, 2013.

[17] D. Bone, C.-C. Lee, and S. S. Narayanan, "A Robust Unsupervised Arousal Rating Framework using Prosody with Cross-Corpora Evaluation." in *INTERSPEECH*, 2012, pp. 1175–1178.

[18] 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.

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

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