AUTOMATIC MEASUREMENT OF AFFECTIVE VALENCE AND AROUSAL IN SPEECH

Meysam Asgari, Géza Kiss, Jan van Santen, Izhak Shafran, and Xubo Song

Center for Spoken Language Understanding
Oregon Health & Science University

ABSTRACT

Methods are proposed for measuring affective valence and arousal in speech. The methods apply support vector regression to prosodic and text features to predict human valence and arousal ratings of three stimulus types: speech, delexicalized speech, and text transcripts. Text features are extracted from transcripts via a lookup table listing per-word valence and arousal values and computing per-utterance statistics from the per-word values. Prediction of arousal ratings of delexicalized speech and of speech from prosodic features was successful, with accuracy levels not far from limits set by the reliability of the human ratings. Prediction of valence for these stimulus types as well as prediction of both dimensions for text stimuli proved more difficult, even though the corresponding human ratings were as reliable. Text based features did add, however, to the accuracy of prediction of valence for speech stimuli. We conclude that arousal of speech can be measured reliably, but not valence, and that improving the latter requires better lexical features.

Index Terms— affect, arousal, valence

1. INTRODUCTION

Computational research on affect has focused more on affect classification (e.g., the 2009–2013 Interspeech Emotion Challenges) than on affect dimensions, specifically the *arousal* (degree of excitement) and *valence* (pleasant vs. unpleasant) dimensions that have deep roots in behavioral research (e.g., [1]). Our interest in dimensional approaches stems from our research on automated measures of behavioral manifestations of *neural underconnectivity* in Autism Spectrum Disorders (ASDs), prompted by recent findings of *functional* underconnectivity as measured via temporal correlations between signals (EEG and fMRI) from different parts of the brain (e.g., [2]) and for *morphological underconnectivity* (e.g., [3]). Neural underconnectivity is expected to have behavioral counterparts via reduced temporal correlations between the behavioral streams in the facial, gestural, and speech modalities. However, in our research, it became clear that discrete affect labels were too unreliable and uninformative (lacking, e.g., affective intensity) to yield stable correlations, while dimensional approaches had the required robustness.

This paper reports on our work on measuring valence and arousal in the speech modality by creating automatic methods that optimally predict "gold standard" human ratings. To better understand the roles of speech contents and speech prosody, we use arousal and valence ratings of three types of stimuli, all based on a corpus of utterances from children: speech, delexicalized speech (prosody), and text transcripts (contents). As a further important methodological detail, we use two approaches that differ in both overall strategy and details in order to demonstrate that results are robust, unlikely to be due to a "lucky" match between data and method.

2. METHODS

2.1. Data

Twenty-eight children ages 8-11 participated: 10 with Typical Development, 11 with ASD, 4 with Specific Language Impairment, and 3 with Unspecified Developmental Delay. Children were trained to re-enact a brief story, pretending to be an actor/actress. Each produced between 19 and 46 utterances (median: 27), 835 utterances in total. The resulting video recordings were split into the (835) per-utterance files. For each of these, five modality-specific versions were produced, for a total of 4175 *stimuli*: *Speech*; *delexicalized speech* (speech rendered unintelligible by methods described in [4]); *face* (no sound, gesture blanked out); *gesture* (no sound, face blanked out); and *text* (produced by human transcribers). Eleven undergraduate students rated arousal and valence of these stimuli on 5-point scales.

Table 1. Average reliability of human group ratings.

Modality	Arousal	Valence
Speech	0.88	0.91
Delexicalized Speech	0.85	0.75
Text	0.75	0.93

Correlations between raters ranged from 0.46 to 0.63, with a median of 0.52. To obtain a measure of the reliability of pooled ratings (which will be our "gold standard"), we one-hundred times randomly split the raters into two sub-

This research was supported by NIH grants R01-DC007129, R21-DC010239, R01-012033, and NSF grant HCC-0905095, to Jan van Santen; and by NIH grant K25-AG033723 and NSF grants RI-0964102 and II-NEW-0958585 to Izhak Shafran. The views herein are those of the authors and do not necessarily reflect the views of the funding agencies.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

groups and computed the (two) sub-group averages for each stimulus. The correlations between these averages ranged from 0.825 to 0.875, and are broken down by affect dimension and modality in Table 1. This establishes that ratings pooled over all 12 raters have adequate reliability.

2.2. Moments+FDA features: features based on statistical moments and Functional Data Analysis

2.2.1. F0 tracking and clean-up

We extracted F0 contours using Snack [5], with 20 ms window length and 10 ms shift, initially tracking F0 with a child-appropriate frequency range of 100-1200 Hz, estimating the frequency range, and tracking the F0 curve again with this range. We calculated the frequency range based on De Looze and Rauzy [6], shown by Keelan et al. [7] to optimize performance of diverse F0 tracking algorithms to almost the same level. For post-processing (which proved helpful only for the Functional Data Analysis approach; see 2.2.3), we automatically identified jumps (halvings and doublings) in the F0 curves and corrected them by multiplying or dividing the F0 values to confirm to the neighboring segments; this step improved the performance significantly. We smoothed the result by applying a median filter with a window size of 5, similarly to Ahmadi and Spanias [8]. We automatically removed silence from the start and end of utterances, based on absence of voicing or very low root mean squared (RMS) values (<40%-ile of the RMS of all unvoiced parts), keeping parts with large intensity (>95%-ile of silence parts) as speech. We transformed both F0 and intensity into the log-domain for feature extraction.

Features considered are based on earlier work on ASD detection from prosodic features ([9]).

2.2.2. Statistical moments based features

We calculated 34 per-utterance statistical moments for F0 and intensity. Non-robust variants were: minimum, mean, maximum, standard deviation, variance, coefficient of variation, skewness, and the logarithm of excess kurtosis. The corresponding robust variants were: 10%-ile, median, 90%-ile, median absolute deviation from the median (MAD), inter-quartile range (IQR), robust coefficient of variation (IQR/median), robust skewness ([10], SK2), and logarithm of robust kurtosis ([10], KR2). We also used the numbers of voiced frames and total speech frames as features, to help the model weigh higher order moments, requiring more analysis frames than lower order moments. The performance does not change significantly if we exclude the location statistics (such as minimum, median, etc.) from the intensity features, which suggests that we need not assume constant microphone-speaker distance.

2.2.3. Functional Data Analysis based features

We used Functional Data Analysis (FDA [11]) utilizing the *fda* R package to characterize the shape of the F0 and intensity curves, calculating the first 10 functional principal component analysis (fPCA) coefficients for F0 and RMS, resulting in 20 features per utterance.

Consider the F0 (or RMS) curve for an utterance y consisting of n time points $t_j, j = 1..n$, and a set of basis functions ϕ_k. We express y as a weighted sum of K basis functions, capturing the important details of the curve:

$$\hat{y} = \sum_{k=1}^{K} c_k \cdot \phi_k. \tag{1}$$

It suffices to know the values of the basis functions at the time points t_j, which we represent as a matrix

$$\Phi = \left[\phi_k(t_j)\right]. \tag{2}$$

We use the least squares error function, assuming Gaussian noise on sample y, and refine the solution by adding to the error function the integral of the second derivative of $\hat{y}(t)$ with respect to t as a roughness penalty, to avoid close fit to the noise. This gives us the regularized least squares solution, with the following estimate for the coefficient vector:

$$\hat{c} = (\Phi^T\Phi + \lambda R)^{-1}\Phi^T y \quad \text{where} \tag{3}$$

$$R = \int D^2\Phi(t) \cdot D^2\Phi^T(t)dt. \tag{4}$$

For a set of N curves, fPCA identifies the first M orthogonal eigenfunctions capturing most variation. We represent curve i as a weighted sum of eigenfunctions ξ_m, with a set of coefficients that maximizes

$$\sum_{i=1}^{N} \int \xi_m(t) \cdot \hat{y}_i(t)dt \tag{5}$$

subject to

$$\int \xi_m^2(t)dt = 1 \text{ and } \int \xi_m(t) \cdot \xi_l(t)dt = 0, \forall l < m. \tag{6}$$

The functional principal component coefficients are given by

$$c_{i,m} = \int \xi_m(t) \cdot (\hat{y}_i(t) - \bar{\hat{y}}(t))\, dt, i = 1..N, m = 1..M \tag{7}$$

for utterance i, eigenfunction m, and mean curve $\bar{\hat{y}}(t)$.

We used 1000 B-splines as basis functions, and calculated the first 10 eigenfunctions and corresponding coefficients for both F0 and RMS on the whole training set, so as to have enough utterances for training a robust model, and calculated the same coefficients for the test utterances using these eigenfunctions, resulting in 20 features per utterance.

2.3. HM features: features based on the Harmonic Model

Alternatively, we extracted prosodic features using the harmonic model [12, 13, 14, 15], a variant of the sinusoidal model where the frequencies of the sinusoidal components

are multiples of F0. Per-utterance features are based on F0 jitter, shimmer, harmonic to noise ratio (HNR), and the energy ratio of the first and second harmonics (H12). In previous work, we found that these features detect voiced segments and estimate F0 more accurately than other algorithms and that they are useful in rating the severity of a subjects Parkinsons disease [16]. Jitter and shimmer (and, of course, F0) have been used previously for emotion detection (e.g., [17]); however, few if any studies on emotion detection have used the harmonic model or have used it for estimating F0, jitter, or shimmer.

2.3.1. Harmonic Model

Briefly, in the harmonic model, the speech samples are represented as follows. Let $\mathbf{y} = [y(t_1), y(t_2), \ldots, y(t_N)]^T$ denote the speech samples in a voiced frame, measured at times $t_1, t_2, \ldots, t_T$. The samples can be represented by the harmonic model with additive noise $\mathbf{n} = [n(t_1), n(t_2), \ldots, n(t_N)]^T$ as follows:

$$\begin{aligned} s(t) &= a_0 + \sum_{h=1}^{H} a_h cos(2\pi f_0 ht) + b_h sin(2\pi f_0 ht) \\ y(t) &= s(t) + n(t) \end{aligned} \quad (8)$$

where H denotes the number of harmonics and $2\pi f_0$ stands for the fundamental angular frequency. Assuming the noise distribution is constant during the frame and is given by $\mathcal{N}(0, \sigma_n^2)$, estimation of the unknown parameters $\Theta = [f_0, a_h, b_h, \sigma_n^2, H]$ can be cast into a maximum likelihood (ML) framework [14]. However, ML estimation of the pitch period may lead to pitch halving and doubling errors. We addressed this problem in our previous work and improved the robustness of the pitch estimates by smoothing the likelihood function [18]. Given the estimates of Θ, we can reconstruct the speech signal for the further analysis.

2.3.2. Jitter and Shimmer

Traditionally, computation of jitter and shimmer assumes that these parameters are constant during the frame. Alternatively, we employ a *harmonic model with varying amplitudes* (HM-VA) [19] that allows within-frame sample-to-sample variations. Our approach for estimating shimmer assumes that speech can be considered as an amplitude modulated (AM) and frequency modulated (FM) signal. To compute shimmer, we first represent the speech waveform using both the HM and the HM-VA. We then use both model parameters for extracting the AM component as a temporal function that represents the amplitude fluctuations [18]. We use its standard deviation over frames as a summary statistic for shimmer.

To estimate jitter, we create a matched filter using a one pitch period long segment from the reconstructed signal and convolve it with the original speech waveform [18]. The distance between the maxima in the convolved signal defines the pitch periods. The perturbation of the period is normalized with respect to the given pitch period and its standard deviation is an estimate of the jitter. Thus, this method allows the computation of jitter within the 25ms analysis window.

2.3.3. Harmonic to noise ratio (HNR)

Once the parameters of the harmonic model are estimated for a frame, the noise can be computed by subtracting the reconstructed signal from the original speech signal. Given the estimated HM parameters for each frame, the HNR and the ratio of the energy in the first and the second harmonics (H12) can be computed as follows.

$$\begin{aligned} c_h &= \sqrt{\sum_{h=1}^{H} a_h^2 + b_h^2} \\ HNR &= \log \sum_{h=1}^{H} c_h^2 - \log \sum_{t=1}^{N} (y(t) - s(t))^2 \\ H12 &= \log c_1 - \log c_2 \end{aligned} \quad (9)$$

2.3.4. Per-Utterance Feature Vectors

We extract 25 ms long frames using a Hanning window with a 10 ms shift. We first detect voiced frames robustly by calculating the likelihood of voicing under the harmonic model. The voicing decision at the segment level is computed by formulating a one-state hidden Markov model (HMM). The state could either be voiced or unvoiced, with likelihood given by the per-frame harmonic model. The transition model consists of a simple zero-mean Gaussian. We compute voicing decision over the utterance using Viterbi alignment. Subsequently, we compute various voicing related features for voiced frames, including F0, HNR, H12, jitter, and shimmer. These pitch-related features are combined with standard features including energy, spectral entropy, and MFCCs. Per-utterance features are computed by applying standard summary statistics such as mean, median, variance, minimum and maximum to the per-frame features, generating a 252-dimensional per-utterance feature vector.

2.4. Features for prediction of text ratings

To extract features from text, we used a published table of valence and arousal ratings by Warriner et al. [20] to tag each word in an utterance with an arousal and a valence rating, and computing their per-utterance mean, standard deviation, minimum, and maximum. For missing words we imputed valence and arousal by randomly drawing 5 words from the table and computing their average. We computed these features from manual and automated transcripts.

For the latter, we built a context-dependent HMM-GMM system with 39-dimensional MFCC features with delta and delta-delta coefficients, using the Kaldi speech recognition toolkit [21]. We used the OGI Kids Speech Corpus, consisting of 27 hours of spontaneous speech from 1100 children, from kindergarten through grade 10 [22]. After cepstral mean

and variance normalization, and LDA, we employed model space adaptation using maximum likelihood linear regression (MLLR). Also, speaker adaptive training (SAT) of the acoustic models were performed by both vocal tract length normalization (VTLN) and feature-space adaptation using feature-space MLLR (fMLLR). A trigram language model was built using the SRILM toolkit ([23]). The WER on a 2-hour test corpus was about 26%.

2.5. Features for prediction of speech ratings

The acoustic features described in subsection 2.3 were concatenated with features extracted from text transcriptions (subsection 2.4) for predicting the speech ratings.

2.6. Prediction Engine

We used Support Vector Regression with an RBF kernel (except that a linear kernel was used for valence in the HM based approach). We used a five-fold cross validation scheme, setting all model parameter using four of the five sets as training set, and using the fifth ones only for reporting the performance estimates below. Parameters of the optimal SVR model were determined on the training set separately for each fold, via grid search and cross-validation. For the Moments+FDA based approach we used the *e1071* R package, and for the harmonic-based approach, we employed the open-source *scikit-learn* toolkit [24].

3. RESULTS

Table 2 indicates that arousal ratings of speech and delexicalized speech can be predicted equally well by the two alternative sets of prosodic features, and are not far from the limits set by the reliability of the human ratings (0.85). In contrast to prediction of valence ratings for delexicalized speech, for speech the HM features (even without text features) performed substantially better than the Moments+FDA features. FDA based features did not add to performance for arousal, but possibly did so for valence, suggesting that curve shape may matter more for valence than for arousal.

Predicting arousal proved challenging for text, with slightly better but still modest results for valence. As expected, ASR transcripts yielded worse results than manual transcripts. Obviously, the word-based approach is simplistic and would miss cases such as "not too bad".

Surprisingly, incorporating text features did not help in the prediction of the arousal rating for speech, and helped very little for valence.

Permutation tests, permuting the predicted values randomly with respect to the observed values, were highly significant ($p<0.0001$), even for the smallest correlation (0.22).

Table 2. Average product-moment correlations between observed and predicted ratings. K is the number of features.

	K	Arousal	Valence
Delexicalized Speech			
FDA	20	0.50	0.31
Moments	34	0.78	0.35
Moments+FDA	54	0.77	0.39
HM	252	0.79	0.42
Text			
Manual Transcript	8	0.36	0.57
ASR Transcript	8	0.26	0.35
Speech			
FDA	20	0.59	0.22
Moments	34	0.83	0.28
Moments+FDA	54	0.83	0.34
HM	252	0.83	0.47
HM + Manual Transcript	260	0.83	0.51
HM + ASR Transcript	260	0.83	0.49
HM + ASR + Moments	294	0.86	0.53

4. DISCUSSION

We conclude the following. First, to the degree that arousal of speech can be reliably measured, prosodic features appear sufficient. Adding text features did not improve performance in the least, despite the fact that prediction of arousal of text, while modest, was statistically significant. A similar trend existed for the human ratings, based on a multiple regression analysis predicting speech ratings from text ratings and delexicalized speech ratings: for arousal, the effect on speech ratings of the latter was four times that of the former, while for valence both had equal effects. Yet, it is clear that better text based features are needed, especially in the light of the very high reliability of the valence ratings (0.93)

Second, the differences in predictive performance as a function of stimulus type (speech, delexicalized speech, text) and dimension (arousal, valence) are not due to differences in reliability of the corresponding human ratings. For example, valence ratings and arousal ratings of speech were equally reliable, yet prediction of the former proved far harder than prediction of the latter. Thus, these differences in predictive accuracy substantially reflect the relevance of the information contained in the respective feature vectors.

Third, for arousal ratings of either speech or delexicalized speech, a ceiling in performance was reached at 34 features. This may be due to the low data-to-parameter ratio being suboptimal for the HM features or to the Moments+FDA features having been more carefully engineered [9].

Our results are near the top of recently reported correlations for arousal predictions (0.62 ([25]) to 0.85 ([26]), but less so for valence (0.29 ([25]) to 0.65 ([27]). These studies used neither delexicalized speech ratings nor text ratings, and are thus inconclusive as to the roles of prosody and lexical contents in the prediction of arousal and valence.

As mentioned, others have used jitter and shimmer for affect classification (e.g., [17]), but not for predicting arousal and valence. The same holds true for usage of Functional Data Analysis (e.g., [28])

5. REFERENCES

[1] R. Abelson and V. Sermat, "Multidimensional scaling of facial expressions.," *Journal of Experimental Psychology*, vol. 63, no. 6, pp. 546, 1962.

[2] S. Schipul, T. Keller, and M. Just, "Inter-regional brain communication and its disturbance in autism," *Frontiers in Systems Neuroscience*, vol. 5, no. 10, pp. 7, 2011.

[3] A. Di Martino, C. Yan, Q. Li, et al., "The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular Psychiatry*, 2013.

[4] A. Kain and J. van Santen, "Frequency-domain delexicalization using surrogate vowels.," in *Interspeech*, 2010, pp. 474–477.

[5] K. Sjölander, "The snack sound toolkit," *KTH*, 2004.

[6] C. De Looze and S. Rauzy, "Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration.," in *Interspeech*, 2009, pp. 2919–2922.

[7] K. Evanini, C. Lai, and K. Zechner, "The importance of optimal parameter setting for pitch extraction," in *Meetings on Acoustics*, 2011, vol. 11, p. 060004.

[8] S. Ahmadi and A.S. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, 1999.

[9] G. Kiss, J. van Santen, E. Prud'hommeaux, and L. Black, "Quantitative analysis of pitch in speech of children with neurodevelopmental disorders.," in *Interspeech*, 2012, pp. 1343–1346.

[10] T. Kim and H. White, "On more robust estimation of skewness and kurtosis," *Finance Research Letters*, vol. 1, no. 1, pp. 56–73, 2004.

[11] J. Ramsay, G. Hooker, and S. Graves, *Functional Data Analysis with R and MATLAB*, 2009.

[12] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," in *Ph.D. dissertation*, 1996.

[13] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *ICASSP*, 2002, vol. 2, pp. 1769–72.

[14] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans on Speech & Audio Processing*, vol. 12, no. 1, pp. 76 – 87, 2004.

[15] M. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[16] Alireza Bayestehtashk, Meysam Asgari, Izhak Shafran, and James McNames, "Fully automated assessment of the severity of parkinson's disease from speech," *Computer Speech & Language*, 2013.

[17] X. Li, J. Tao, M. Johnson, et al., "Stress and emotion classification using jitter and shimmer features," in *ICASSP*, 2007, vol. 4, pp. IV–1081.

[18] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders," in *Interspeech*, 2013.

[19] M. Asgari, I. Shafran, and A. Bayestehtashk, "Robust detection of voiced segments in samples of everyday conversations using unsupervised hmms," in *IEEE SLT*, 2012, pp. 438–442.

[20] A. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior Research Methods*, pp. 1–17, 2013.

[21] D. Povey, A. Ghoshal, G. Boulianne, et al., "The kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.

[22] K. Shobaki, J. Hosom, and R. Cole, "The OGI kids' speech corpus and recognizers," in *ICSLP*, 2000.

[23] A. Stolcke et al., "Srilm-an extensible language modeling toolkit.," in *Interspeech*, 2002, pp. 1618–1621.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] K. Truong, D. Van Leeuwen, and F. De Jong, "Speech-based recognition of self-reported and observed emotion in a dimensional space," *Speech Communication*, vol. 54, no. 9, pp. 1049–1063, 2012.

[26] D. Wu, T. Parsons, E. Mower, and S. Narayanan, "Speech emotion estimation in 3d space," in *International Conference on Multimedia and Expo*. IEEE, 2010, pp. 737–742.

[27] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[28] J. Arias, C. Busso, and N. Yoma, "Energy and f0 contour modeling with functional data analysis for emotional speech detection," in *Interspeech*, 2013.