"Honey, I Learned to Talk": Multimodal Fusion for Behavior Analysis

Shao-Yen Tseng*

University of Southern California
Los Angeles, California
shaoyent@usc.edu

Brian Baucom

The University of Utah
Salt Lake City, Utah
brian.baucom@utah.edu

Haoqi Li*

University of Southern California
Los Angeles, California
haoqili@usc.edu

Panayiotis Georgiou

University of Southern California
Los Angeles, California
georgiou@sipi.usc.edu

ABSTRACT

In this work we analyze the importance of lexical and acoustic modalities in behavioral expression and perception. We demonstrate that this importance relates to the amount of therapy, and hence communication training, that a person received. It also exhibits some relationship to gender. We proceed to provide an analysis on couple therapy data by splitting the data into clusters based on gender or stage in therapy. Our analysis demonstrates the significant difference between optimal modality weights per cluster and relationship to therapy stage. Given this finding we propose the use of communication-skill aware fusion models to account for these differences in modality importance. The fusion models operate on partitions of the data according to the gender of the speaker or the therapy stage of the couple. We show that while most multimodal fusion methods can improve mean absolute error of behavioral estimates, the best results are given by a model that considers the degree of communication training among the interlocutors.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by regression**;

KEYWORDS

Behavior analysis, multimodal fusion, expert fusion

ACM Reference Format:

Shao-Yen Tseng*, Haoqi Li*, Brian Baucom, and Panayiotis Georgiou. 2018. "Honey, I Learned to Talk": Multimodal Fusion for Behavior Analysis. In *2018 International Conference on Multimodal Interaction (ICMI '18)*, October 16–20, 2018, Boulder, CO, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3242969.3242996>

*These authors contributed equally to this work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMI '18, October 16–20, 2018, Boulder, CO, USA

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5692-3/18/10...\$15.00

<https://doi.org/10.1145/3242969.3242996>

1 INTRODUCTION

Observational studies of human behavior are important in a wide range of domains from social behavior analysis to health care treatment. Based on the study of human behaviors, domain experts can give diagnostic evaluations and design further treatments. For instance, in the field of psychology, psychologists use observable interaction cues to improve the treatment plan. However, considering the huge amount of manual processing work and resources involved in traditional observational practice, it is hard to employ observational studies widely. Behavioral Signal Processing (BSP) [10, 17] is developing algorithms and systems to combine machine learning, signal processing and domain-specific knowledge to quantitatively understand and model human behavior especially in speech and language-based affective behaviors and social behaviors.

Over the last few years, BSP employed lexical [5, 9], acoustic [14, 19] and visual information [25] to analyze and model human-centered behavior in domains like Couple Therapy [14, 15, 19, 23], Addiction [5, 26, 27], Suicide [18], and Cancer [20]. However, building an automatic behavior analysis system is still not an easy task due to several aspects that we expand on below.

One challenging part of BSP relates to the complex and multi-faceted aspect of human behavior. These manifest through an intricate interplay between brain and body [17], thus the evaluation or annotation of behaviors can be regarded as a highly complex non-linear process. Moreover, humans can convey behavioral information via multiple modalities, including the visual, lexical and acoustic channels, which may not always be congruent. Although using single modality information can automatically classify or estimate behaviors [14, 23] quite accurately, there is still room for improvements. Considering only one modality can result in information loss and potentially lower accuracy of the behavior recognition task. Further employing multiple modalities may capture information due to the relative importance and (in)congruency of the modalities. There are existing efforts in exploring fusion in affect or emotion recognition domains [4, 13, 16, 21, 28] as well as work in the couples therapy domain [2].

Multimodal fusion has been shown to improve performance in many domains. However one of the major challenges in multimodal fusion is accounting for modality saliency, where the amount of relevant information in each modality varies. To address this issue for example [7] proposed a modality saliency detection framework

for biased fusion of infrared and visible light images. In [11] a mixture-of-experts fusion method was proposed which dynamically combined information from audio and video for predicting emotion in movies.

In studies of human behavior, the issue of modality importance is especially prominent since behavior expression varies from person to person, and even for the same person depending on state of mind (internal) or external influences. Further, the amount of information encoded in each multimodal stream is itself meaningful. For example speakers that know how to communicate well can better receive and express information reducing the degree of Sentiment Override [8]. Participants in the analyzed study participated in traditional or integrative behavioral couple therapy (TBCT or IBCT) and their participation in the study increased their positivity and problem solving [22]. As we will show in this work, this also correlates with an increased degree of information carried through content (lexical) as opposed to acoustics.

In this paper, we first analyze the data and show that the degree of Behavioral Couple Therapy (IBCT or TBCT), that we will call "Therapy Stage" is a factor on the importance of each modality, even more so than gender. We then show that by partitioning the data along along these meaningful clusters we can observe specific trends towards modality importance. We thus proceed to propose the use of expert multimodal fusion models for behavior recognition which operate on knowledge-based partitions of the data. We hypothesize that different fusion models can be trained to attend to differing modality biases along each cluster. We evaluate our assumption by estimating the real-valued behavior ratings from acoustic and lexical modalities.

2 DATASET AND FEATURE EXTRACTION

2.1 Couple Therapy Data Corpus

In this paper, we use behavioral data collected by the UCLA/UW Couple Therapy Research Project [6], where 134 couples participated in video-taped sessions while discussing marital issues. During each session, couples discussed one self-selected relationship-related topic for around 10 minutes. Couples were defined as a *Wife* and *Husband* pair in the dataset resulting in a **Gender** partition. Session recordings took place before therapy began (*pre*), after 26 weeks (*26wk*), and after 2 years of therapy (*2yr*). We refer to these recording timings as the **Therapy Stages**. For every recorded session, each participant was separately rated on a scale of 1 to 9 for multiple behaviors by well-trained human annotators based on the Couples Interaction and Social Support Rating Systems [12]. In this paper, we focused on three behavior codes: *Negativity*, *Blame*, and *Acceptance*. More details of this corpus can be found in [3].

2.2 Preprocessing and feature extraction

In our experiment, we utilized those sessions with SNR above 5dB, applied speech activity detection and performed speaker diarization to identify speaker regions between *Husband* and *Wife*. We extracted the same Low Level-Descriptors (LLDs) features and performed same normalization processing as in [14], after which we employed a frame window size of 10s to calculate the frame level statistical functions as the final acoustic feature set. For our lexical analysis we used human transcriptions. Existing work (e.g. [2, 9]

has shown the ability to do this analysis from automatically generated transcripts, but this isn't addressed in this work due to the multimodal focus and how it links to communication patterns.

To train all unimodal systems as well as the fusion models in Sections 3 and 5, we used a leave-one-couple-out cross-validation scheme. That is, for each fold we separated all data from one couple as test data and used the remaining as training data.

3 UNIMODAL SYSTEMS

We first implement and analyze the performance of two behavior systems on lexical and acoustic modalities.

3.1 Acoustic Approach

We employed the SD-DNN framework proposed in [14] as our acoustic-feature based behavior recognition system. The SD-DNN separated the DNN training into two stages: First, based on the feature type, the whole set of acoustic features was split into subsets and multiple smaller distinct DNN classifiers were trained on each feature subset. In the second stage, all hidden layers of these classifiers were frozen and pre-trained weights were integrated to become parts of a deeper network for further training. This overall system allows the full network to consider information from all feature sets while limiting the number of trainable parameters during each training stage to achieve convergence with a limited amount of data. Through this framework, we obtain sliding window based behavior scores as frame-level behavior scores.

3.2 Lexical Approach

For the lexical system we implemented the behavior scoring method proposed in [23]. The authors proposed the use of conversational sentence embeddings trained in an unsupervised manner using contextual information and the notion of behavioral short-term stationarity. The sentence embeddings were trained to connect two consecutive utterances of movies using a sequence-to-sequence model, and the embedding connecting the two is employed as features. A subsequent neural network using LSTMs was then applied to the sentence embeddings on the in-domain dataset to estimate final behavior scores of each sentence. We refer to these sentence-level behavior scores as lexical frame-level scores.

3.3 Session Ratings Estimation

The unimodal systems described above generate predictions of behavior scores at each frame. The frame size, while different for lexical and acoustic streams, encompasses short subsets of the session. To estimate behavior ratings for the entire session, we applied a Support Vector Regressor (SVR) using RBF-kernel on statistics of the frame-level predictions to map back to the real-valued human ratings. The statistics we used were the *median*, *mean*, and *variance* of frame-level scores in each session.

4 ANALYSIS OF MODALITY IMPORTANCE

4.1 Optimal Weighting

In the first step, we analyzed the importance of each modality by calculating an optimal weighting of acoustic and lexical scores for estimating the final rating. This is implemented by a fitting a linear least mean squares (LMS) regression model on each session.

Figure 1: Box plot of the optimal lexical weight per session grouped by different partition methods for *Negativity*.

Suppose S_A and S_L are session ratings generated through acoustic system and lexical system respectively, w_1 and w_2 are the corresponding weights, and R is the true rating value, then the optimal weights are derived by minimizing $\|w_1 S_A + w_2 S_L - R\|^2$ subject to the constraints $w_1, w_2 \geq 0$ and $w_1 + w_2 = 1$. A value of 1 for lexical weight would imply that the lexical modality is solely used to predict the true rating in the LMS model and no extra information is given by the acoustic modality.

To visualize intercluster differences, we grouped the results based on Gender or Therapy Stage and collected statistics of the weights assigned to the lexical modality. The box plot [24] of these statistics are shown in Figure 1. These findings are consistent with the concept that as couples are improving in their communication skills through therapy, the effects of noncontingent responses are reduced. This manifests as an increasing importance of lexical information at 26-week and 2-year follow-ups, suggesting that behaviors are more strongly tied to the lexical content after treatment when an increasing majority of couples are no longer clinically distressed.

4.2 Analysis per partition

As further analysis, we also trained separate LMS regression models on each subset of the data partitioned by Gender or Therapy Stage. We then analyzed the histogram of weight values assigned to the lexical modality. The histograms of lexical weights for Gender based and Therapy Stage based separation of *Blame* and *Negativity* are shown in Figure 2.

We observed that there was a difference in the values of lexical weights in behaviors of *Blame* and *Negativity* for both clustering methods. For *Acceptance* there was no clear separation in either method. One explanation is the fact that the positive behavior (i.e. *Acceptance*) under investigation is not as affected by BCT. Positive aspects are already expressed in a constructive manner. However, negative behaviors (i.e. *Negativity*, *Blame*) are due to BCT better expressed with increased therapy exposure [22].

5 MULTIMODAL FUSION

5.1 Knowledge-Driven Expert Fusion

Motivated by the analysis in Section 4, we propose a knowledge-driven fusion framework based on data partitioning using **Gender** or **Therapy Stage** information. Given a separation method (Gender or Therapy Stage), the data is split into multiple partitions. Then in each partition an SVR fusion model is trained to estimate true ratings from session-level scores produced by the two modalities. This yields fusion models that are optimized for each partition. During testing, the fusion score is obtained by selecting the output

Table 1: Per-Gender Statistics of behavior

Statistic (%)	<i>Negativity</i>	<i>Blame</i>	<i>Acceptance</i>
Lexical system			
Husband accuracy	82.1	80.0	83.6
Wife accuracy	90.0	89.3	87.1
Overall accuracy	86.1	84.6	85.4
Acoustic system			
Husband accuracy	74.3	65.7	70.0
Wife accuracy	77.9	81.4	75.7
Overall accuracy	76.1	73.6	72.9

from the SVR model which corresponds to the partition the test data belongs to. To simulate scenarios where Therapy Stage information is unavailable we also trained a support vector machine (SVM) to automatically predict the Therapy Stage given unimodal session-level scores estimates. The output of this SVM is then used to select the corresponding expert fusion model. The framework of the fusion system is shown in Figure 3.

5.2 Classical Multimodal Fusion

We compare our results to classical fusion methods such as statistics-based and decision-level fusion models.

5.2.1 Statistics-based Fusion. Since the modalities have different frame-level resolution we instead combine the statistics of the modality features for fusion. An SVR is then trained on the concatenated statistics to obtain a mapping to the true rating. The use of statistic features also eliminates the need for alignments between acoustic and lexical outputs.

5.2.2 Decision-level Fusion. In this scheme an additional SVR was trained to estimate true ratings from the unimodal score estimates given by an SVR in each modality.

6 EVALUATION RESULTS

6.1 Classification Performance

Table 1 shows the binary classification accuracy of the lexical and acoustic systems. For all three behaviors, the overall accuracy of the acoustic system is lower than that of the lexical system. From an empirical perspective, the issue of gender differences in recognizing behaviors is still a topic of debate. To investigate this gender difference, we compare the classification accuracy of *Husband* and *Wife* respectively. Although we applied feature normalization per gender per session described in Section 2.2, in which we try to eliminate the speaker characteristics information from different speakers, and try to train the classification model to only focus on behavior related information, the results still show distinguished differences between genders. The behavior classification accuracy of *Wife* is consistently higher than that of *Husband*, which means it is easier to obtain correct recognized behaviors for female than male. This result can be one evidence of gender difference in behavior expression as also shown in emotion recognition tasks such as [1].

6.2 Multimodal Behavior Ratings

As we are predicting real-valued behavior ratings, we evaluated the performance using the mean absolute error (MAE) of the predicted ratings to the average human rating value. The results of the various

Table 2: Mean Absolute Error of Fusion Ratings (1-9)

Method	<i>Neg.</i>	<i>Blame</i>	<i>Accept.</i>	All
Baseline lexical [23]	1.37	1.70	1.23	1.43
Baseline acoustic [14]	2.51	2.73	1.96	2.40
Statistics-based fusion	1.40	1.54	1.25	1.40
Decision-level fusion	1.31	1.40	1.21	1.31
Gender based fusion	1.31	1.41	1.22	1.31
Therapy Stage fusion	1.22	1.28	1.18	1.23
Therapy Stage fusion (Auto)	1.25	1.28	1.14	1.22

fusion methods are shown in Table 2. The MAE represents the average deviation from the true rating and is an intuitive measure of the accuracy of our estimator. In our experiments the decision-level and all methods of knowledge-driven expert fusion improved the rating accuracy for all behaviors, while statistics-based fusion only gave minor improvement for *Blame*. Decision-level fusion performed very similarly to Gender based fusion with relative improvements from a unimodal lexical system of 3%, 18%, and 2% for *Negativity*, *Blame*, and *Acceptance* respectively. Therapy Stage based fusion, as shown in the second last row of the table, performed the best in all behaviors with relative improvements of 11%, 25%, and 5% for *Negativity*, *Blame*, and *Acceptance* respectively. The last row of the table shows the results of SVM based automatic Therapy Stage identification which has similar performance to oracle expert selection.

7 DISCUSSION

Our results validate our assumption that the importance of different modalities for behavioral annotation not only varies between genders, but also with the degree of therapy, TBCT or IBCT. This does not only support the creation of improved fusion algorithms but most importantly reinforces the validity of TBCT and IBCT methods.

Figure 3: Architecture of expert multimodal fusion

The finding that lexical information is increasingly important with the degree of BCT is also intuitive: the purpose of therapy is in itself to improve the behavior of participants, enable them to better participate in discussions, better follow their partner and reduce negative sentiment override behaviors. All these promote better expression and communication by the partners, which is similar to the conclusion from work [22].

Another interesting observation is that while there is evidence of different modality saliency for each gender, the performance of Gender based fusion is similar to that of a single fusion model using SVR. This result could be attributed to the higher dimensionality of the SVR which can overcome differences in Gender based modality biases. However, we still observed improvements in performance through the use of Therapy Stage based fusion methods using SVR.

8 CONCLUSIONS AND FUTURE WORK

In this work we applied and compared various fusion methods to estimate real-valued behavior ratings using outputs from unimodal acoustic and lexical systems. We proposed the use of multiple expert fusion models and applied the framework on a Gender based and Therapy Stage based partition scheme as a proof of concept. We showed that late fusion models estimated behavior ratings more accurately than individual acoustic or lexical systems. In particular, we found that by separating the data based on Therapy Stage and training expert fusion models for each partition we were able to achieve more accurate behavior annotations. To obtain insight for this result we also performed an analysis of modality saliency for the different partition schemes using a linear LMS regression model.

Evidently, the expression of behavior through different modalities varies across individuals and degree of BCT exposure. To obtain accurate estimates of behavior, it is important to account for modality saliency when implementing multimodal fusion. In practice it is very difficult to obtain the modality weights a priori, however, we demonstrate our assumption using knowledge-based partitions. In future work it would be beneficial to explore unsupervised ways of determining modality saliency through further analysis of the frame-level behavior score trajectories.

ACKNOWLEDGMENTS

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Research Program under Award No. W81XWH-15-1-0632.

REFERENCES

- [1] Kaat Alaerts, Evelien Nackaerts, Pieter Meyns, Stephan P Swinnen, and Nicole Wenderoth. 2011. Action and emotion recognition from point light displays: an investigation of gender differences. *PLoS one* 6, 6 (2011), e20989.
- [2] Matthew P Black, Panayiotis G Georgiou, Athanasios Katsamanis, Brian R Baucom, and Shrikanth Narayanan. 2011. "You made me do it": Classification of Blame in Married Couples' Interactions by Fusing Automatically Derived Speech and Language Information. In *Twelfth Annual Conference of the International Speech Communication Association*.
- [3] Matthew P Black, Athanasios Katsamanis, Brian R Baucom, Chi-Chun Lee, Adam C Lammert, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2013. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech communication* 55, 1 (2013), 1–21.
- [4] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 205–211.
- [5] Dogan Can, Panayiotis G. Georgiou, David Atkins, and Shrikanth S. Narayanan. 2012. A Case Study: Detecting Counselor Reflections in Psychotherapy for Addictions using Linguistic Features. In *Proceedings of Interspeech*.
- [6] Andrew Christensen, David C Atkins, Sara Berns, Jennifer Wheeler, Donald H Baucom, and Lorelei E Simpson. 2004. Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. *Journal of consulting and clinical psychology* 72, 2 (2004), 176.
- [7] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia* 15, 7 (2013), 1553–1568.
- [8] Frank D Fincham, Philip C Garner, Susan Gano-Phillips, and Lori N Osborne. 1995. Preinteraction expectations, marital satisfaction, and accessibility: A new look at sentiment override. *Journal of Family Psychology* 9, 1 (1995), 3.
- [9] Panayiotis G Georgiou, Matthew P Black, Adam C Lammert, Brian R Baucom, and Shrikanth S Narayanan. 2011. "That's Aggravating, Very Aggravating": Is It Possible to Classify Behaviors in Couple Interactions Using Automatically Derived Lexical Features?. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 87–96.
- [10] Panayiotis G Georgiou, Matthew P Black, and Shrikanth S Narayanan. 2011. Behavioral signal processing for understanding (distressed) dyadic interactions: some recent developments. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. ACM, 7–12.
- [11] Ankit Goyal, Naveen Kumar, Tanaya Guha, and Shrikanth S Narayanan. 2016. A multimodal mixture-of-experts model for dynamic emotion prediction in movies. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2822–2826.
- [12] C Heavey, D Gill, and A Christensen. 2002. Couples interaction rating system 2 (CIRS2). *University of California, Los Angeles* 7 (2002).
- [13] Chul Min Lee, Shrikanth S Narayanan, and Roberto Pieraccini. 2002. Combining acoustic and language information for emotion recognition. In *Seventh International Conference on Spoken Language Processing*.
- [14] Haoqi Li, Brian Baucom, and Panayiotis Georgiou. 2016. Sparsely Connected and Disjointly Trained Deep Neural Networks for Low Resource Behavioral Annotation: Acoustic Classification in Couples Therapy. *Interspeech 2016* (2016), 1407–1411.
- [15] Haoqi Li, Brian Baucom, and Panayiotis Georgiou. 2017. Unsupervised latent behavior manifold learning from acoustic features: Audio2behavior. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 5620–5624.
- [16] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. 2010. Decision level combination of multiple modalities for recognition and analysis of emotional expression. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2462–2465.
- [17] Shrikanth Narayanan and Panayiotis G Georgiou. 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proc. IEEE* 101, 5 (2013), 1203–1233.
- [18] Md Nasir, Brian Baucom, Craig J Bryan, Shrikanth Narayanan, and Panayiotis Georgiou. 2017. Complexity in speech and its relation to emotional bond in therapist-patient interactions during suicide risk assessment interviews. In *Proceedings of Interspeech*.
- [19] Md Nasir, Wei Xia, Bo Xiao, Brian Baucom, Shrikanth S Narayanan, and Panayiotis G Georgiou. 2015. Still together?: The role of acoustic features in predicting marital outcome. In *Sixteenth Annual Conference of the International Speech Communication Association*.
- [20] Maija Reblin, Richard E Heyman, Lee Ellington, Brian RW Baucom, Panayiotis G Georgiou, and Susan T Vadaparampil. 2018. Everyday couples communication research: Overcoming methodological barriers with technology. *Patient education and counseling* 101, 3 (2018), 551–556.
- [21] Björn Schuller, Ronald Müller, Manfred Lang, and Gerhard Rigoll. 2005. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Ninth European Conference on Speech Communication and Technology*.
- [22] Mia Sevier, Kathleen Eldridge, Janice Jones, Brian D. Doss, and Andrew Christensen. 2008. Observed Communication and Associations With Satisfaction During Traditional and Integrative Behavioral Couple Therapy. *Behavior Therapy* 39, 2 (2008), 137 – 150. <https://doi.org/10.1016/j.beth.2007.06.001>
- [23] Shao-Yen Tseng, Brian Baucom, and Panayiotis Georgiou. 2017. Approaching human performance in behavior estimation in couples therapy using deep sentence embeddings. *Proceedings of Interspeech* (2017).
- [24] John W Tukey. 1977. *Exploratory data analysis*. Vol. 2. Reading, Mass.
- [25] Bo Xiao, Panayiotis Georgiou, Brian Baucom, and Shrikanth S Narayanan. 2015. Head motion modeling for human behavior analysis in dyadic interaction. *IEEE transactions on multimedia* 17, 7 (2015), 1107–1119.
- [26] Bo Xiao, Chewei Huang, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan. 2016. A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Computer Science* 2 (2016), e59.
- [27] Bo Xiao, Zac E Imel, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2015. "Rate My Therapist": Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. *PLoS one* 10, 12 (2015), e0143055.
- [28] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* 31, 1 (2009), 39–58.