"You made me do it": Classification of Blame in Married Couples' Interactions by Fusing Automatically Derived Speech and Language Information

Matthew P. Black[1], *Panayiotis G. Georgiou*[1], *Athanasios Katsamanis*[1], *Brian R. Baucom*[2], *Shrikanth S. Narayanan*[1,2]

[1]Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA
[2]Department of Psychology, University of Southern California, Los Angeles, CA, USA
http://sail.usc.edu

Abstract

One of the goals of behavioral signal processing is the automatic prediction of relevant high-level human behaviors from complex, realistic interactions. In this work, we analyze dyadic discussions of married couples and try to classify extreme instances (low/high) of blame expressed from one spouse to another. Since blame can be conveyed through various communicative channels (e.g., speech, language, gestures), we compare two different classification methods in this paper. The first classifier is trained with the conventional static acoustic features and models "how" the spouses spoke. The second is a novel automatic speech recognition-derived classifier, which models "what" the spouses said. We get the best classification performance (82% accuracy) by exploiting the complementarity of these acoustic and lexical information sources through score-level fusion of the two classification methods.

Index Terms: behavioral signal processing (BSP), couple therapy, blame, acoustic features, lexical features, fusion

1. Introduction

Traditional speech and language processing research focuses on detecting more objective human processes. For example, automatic speech recognition (ASR) attempts to map speech signals to written language, and there have been steady improvements over the past decades to each of the components of the recognition process (e.g., acoustic/language/dialog modeling).

In recent years, there have been increased research efforts on modeling more abstract human behaviors/states, such as affect/emotions [1–3] and other paralinguistic phenomena like intent [4] and likability [5]. This has led to the emergence of behavioral signal processing (BSP) [6,7], whose primary goal is to quantify and recognize complex human behaviors in naturally occurring interaction settings, especially those that are relevant to psychology and health-related research.

This paper builds upon our early BSP work [6], in which we analyzed a large corpus of married couples discussing a problem in their relationship. Each spouse was manually rated with a number of session-level behavior codes (e.g., level of blame), as guided by expert psychologists in the domain [8]. We showed that we could automatically classify extreme instances (low/high) for a subset of these codes significantly better than chance using static classifiers trained on functionals (e.g., mean) of frame-level acoustic low-level descriptors (e.g., f_0) [6]. While these initial results were promising, there was still a performance gap in reaching human expert-like ratings.

In this work, our goal is to reduce this performance gap. We concentrated on predicting a single code, the spouse's "level of blame." Blame is particularly relevant for this type of discussion and is oftentimes targeted in couple therapy, since blaming behavior can lead to an escalation of negative affect and resentment between the spouses [9]. Automatically detecting a spouse's level of blame from objective signal-based cues could provide psychologists an alternative and potentially more consistent procedure to quantitatively code human interaction data; it could also provide insight into the production/perception relationship of this human behavior.

We found blame to be one of the more challenging codes to predict using the static acoustic classification method applied in [6]. We hypothesized that this method, which employed session-level statistics of acoustic features, was not modeling the dynamics of the interaction adequately, and furthermore was ignoring important lexical cues regarding blame. The coding manual used to rate blame says that, "explicit blaming statements (e.g., 'you made me do it') warrant a high blame score [8]," and the acoustic features we extracted in [6] were not able to capture these types of spoken phenomena.

In this paper, we propose a number of extensions to our previous work. First, we improve upon the static acoustic classifier by extracting additional hierarchical features that better capture moment-to-moment changes in the interaction. Second, we introduce an ASR-derived classification method that incorporates lexical information through the use of two competitive maximum likelihood language models (one trained on "low blame" text and the other trained on "high blame" text). We show that even with noisy ASR, this method is able to capture discriminative aspects of blame behaviors. Moreover, we show that we can attain the highest classification performance by combining the complementary acoustic and language information sources through score-level fusion of the two classification methods. As part of this work, we also provide an upper bound on performance by running an oracle experiment for the case when we have access to perfect word-level transcriptions.

Section 2 explains the corpus and the classification set-up. Section 3 describes the acoustic features we extracted, and Section 4 explains the various classification methods. We report our results and provide a discussion in Section 5, and we offer our conclusions and plans for future work in Section 6.

2. Corpus

2.1. Background

We used data collected as part of the largest longitudinal, randomized control trial of psychotherapy for severely and stably distressed couples [10]. The corpus consists of 569 ten-minute wife-husband interactions (from 117 married couples), in which they discussed a problem in their relationship. Each spouse's overall level of blame was coded on a 9-point scale by multiple

trained evaluators using a standardized coding manual [8].

The data consist of a single channel of far-field audio (with variable noise conditions across the sessions) and corresponding word-level transcription. The average estimated signal-to-noise ratio (SNR) for the sessions ranged from -1 dB to 26 dB, based on a voice activity detector (VAD) trained on a held-out session [11]. In this paper, we ignored all sessions with an average SNR less than 5 dB. The word-level transcriptions were chronological and included speaker labels (wife or husband), but they did not contain any timing information. We provide a more detailed description of the corpus in [6].

2.2. Speaker Segmentation

For data involving multi-person interactions, manually segmenting by speaker is a common pre-processing step. We took a unique manual/automatic "hybrid" approach to speaker segmentation by exploiting the available transcriptions. Using freely-available software we developed, *SailAlign* [12], we split each session into wife/husband/unknown speaker-homogeneous regions using a recursive speech-text alignment procedure. The wife/husband regions can be thought of as "pseudo speaker turns," since portions of the actual turn may not have been successfully segmented using this technique.

This hybrid speaker segmentation method has two main advantages over manual segmentation: 1) the resulting segmentation is more representative of the hypotheses that would be attained by fully automatic methods (i.e., both methods break down under challenging conditions like overlapped speech and noisy portions of audio), and 2) the hybrid method provides an objective way to reject data that is too noisy to process for acoustic pattern recognition tasks. For this corpus, we ignored all sessions for which we could not automatically align at least 55% of both the husband's and wife's transcribed words.

Of the 569 sessions, 372 met both the 5 dB SNR and 55% speaker segmentation thresholds, which left us 62.8 hours of data across 104 unique couples. For the remainder of the paper (with the exception of the oracle experiments), we treat the speaker segmentation hypotheses as if they were generated by a fully automatic system that only relied upon the audio signal, i.e., we ignore the fact that we have knowledge of the lexical content of each segmented wife/husband region. We do this to simulate more realistic test conditions, where we would normally not have access to what each spouse said.

2.3. Classification Set-up

Since our goal was to automatically separate extreme instances of blame exhibited by the spouses, we used a binary classification set-up similar to the ones used in [4, 6] and partitioned the data into two classes: *high* blame and *low* blame. The high blame partition consisted of the 70 sessions (approximately 20% of the 372 sessions) with the highest average blame score for the wife and the 70 sessions with the highest average blame score for the husband. The low blame partitions consisted of the 140 sessions with the lowest average blame score: 70 for the wife and 70 for the husband. The blame scores for the two classes ranged from 1.0-1.5 for low blame and 5.0-9.0 for high blame, so they were separable to the human evaluators.

Whereas in [6] we trained gender-specific models, in this paper we chose to train gender-independent models, thus effectively doubling the amount of training data. We chose *accuracy* to be the performance metric, defined as the percentage of correctly classified test sessions (out of 280); baseline chance accuracy is 50%. To ensure that the reported results were not overstated, we used leave-one-couple-out cross-validation to separate training and test data, and we optimized all classifier parameters at each train/test fold by using leave-two-couples-out cross-validation on the training data. Therefore, there was no "contamination" of the test couple during the training stages.

LLD	speech/non-speech, f_0, intensity, 15 MFCCs, 8 MFBs, jitter, jitter-of-jitter, shimmer
Functional	mean*, median*, standard deviation*, minimum*, maximum*, range*, skewness, kurtosis, min/max positions, lower quartile, upper quartile, interquartile range, linear approximation slope coeff.

Table 1: *A list of the acoustic low-level descriptors (LLDs) and static functionals we used; the six "basic" functionals are starred (*).*

3. Acoustic Feature Extraction

3.1. Low-Level Descriptors

Table 1 lists the various acoustic low-level descriptors (LLDs) we extracted across each session using standard short-time speech signal processing techniques; we chose these LLDs, based on our previous work [6] and related work in emotion recognition [1–3]. The speech/non-speech LLD was estimated using the VAD [11] as described in Section 2.1. The prosodic LLDs (f_0, intensity) were extracted with Praat [13] and subsequently filtered and normalized as described in [6]. The remaining spectral and voice quality LLDs were extracted with openSMILE [14] using the parameter settings proposed in [15].

Section 3.2 describes how we generated the final static acoustic features from these LLDs. The lexical classification method we implemented (Section 4.2) is based on ASR within the hidden Markov model framework. We used the standard frame-level 39-dimensional vector: the first 13 mean-subtracted Mel-frequency cepstral coefficients (MFCCs) and their first-order derivative (Δ) and acceleration ($\Delta\Delta$) coefficients.

3.2. Static Acoustic Features

We took an overgenerative approach to producing the static acoustic features. The features were static *functionals* (Table 1), computed for each *LLD* (Table 1) across five different *speaker regions* and at six different *temporal granularities*.

The five speaker regions were: 1) where the rated spouse was speaking, based on the speaker segmentation in Section 2.2; 2) where the partner of the rated spouse was speaking; 3) where the wife was speaking, regardless of who was being rated; 4) where the husband was speaking; 5) the entire session, regardless of who was speaking and who was being rated.

The six temporal granularities included one set of *global* features, in which functionals were computed across the entire session (for each LLD and speaker region), and five sets of *hierarchical* features, based on [16]. The five hierarchical feature sets were computed by first splitting the LLD/speaker region into disjoint windows of durations: 0.1s, 0.5s, 1s, 5s, 10s. We then computed the 14 functionals listed in Table 1 for each of these windows, producing 14 vectors of functional values for the entire session. Finally, we generated the hierarchical features by computing the six "basic" functionals (Table 1) across each of these vectors. Because of the windowing technique used, these hierarchical features hopefully capture some of the moment-to-moment changes that occur within the interaction; note that we only computed global features in [6].

After removing static features with zero standard deviation, there were about 53100 features at each cross-validation fold.

4. Classification & Fusion Methods

Sections 4.1-4.4 describe the static acoustic, ASR-derived lexical, oracle lexical, and fusion classifiers, respectively. Figure 1

is a block diagram for the signal-driven classification methods.

4.1. Static Acoustic Classifier

The static acoustic classifier finds a mapping from the high-dimensional static acoustic feature space, which represent various properties of the spouses' speech, to the binary blame class labels. We used the support vector machine (SVM) implementation in LIBSVM [17]. Since there were orders of magnitude more features (50,000+) than instances (280), we used a linear kernel. All features were z-normalized by subtracting the mean value in the training data and dividing by the standard deviation.

4.2. ASR-derived Lexical Classifier

The main problem in using ASR to derive lexical information is the resulting "noisy" word hypotheses, due to numerous factors (e.g., noisy audio, mismatched acoustic/language models). We partially circumvented this noisy ASR problem by implementing an ASR-derived lexical classifier, which incorporated differences in language use between *low* and *high* blame spouses via competitive language models. We derive the equation for this classifier in Equations 1-7, based on [18].

$$[B^*, W^*] = \underset{B,W}{\operatorname{argmax}}\ P(B, W|O),\ \ B \in \{-1, 1\} \tag{1}$$

$$\approx \underset{B,W}{\operatorname{argmax}} \prod_t P(B, W_t|O_t) \tag{2}$$

$$= \underset{B,W}{\operatorname{argmax}} \prod_t P(O_t|W_t, B)P(W_t|B) \tag{3}$$

$$\approx \underset{B,W}{\operatorname{argmax}} \prod_t P(O_t|W_t)\tilde{P}(W_t|B) \tag{4}$$

Equation 1 states to choose the most probable blame class $B \in \{-1, 1\}$ (low/high blame) and most likely word sequence W, given the acoustic observations O of the rated spouse's speech; we disregard the speech regions of the rated spouse's partner for this classification implementation. For computational reasons, we assume in Equation 2 that each speaker turn is independent, and we denote the acoustic observations and word sequence of turn t as O_t and W_t, respectively. We attain Equation 3 by applying Bayes' theorem and dropping the B prior, since both blame classes are equally represented in our experiments. Equation 3 is a variation of the fundamental equation for ASR, where $P(O_t|W_t, B)$ corresponds to the "blame class"-specific acoustic model (AM), and $P(W_t|B)$ corresponds to the "blame class"-specific language model (LM).

For this initial work, we did not train AMs for both blame classes and instead used generic AMs; thus, we assumed that the acoustic observations were independent from B, as shown in Equation 4. We trained the "blame class"-specific LMs using the transcriptions of spouses in the training data at each cross-validation fold: a "high blame" LM on the text from spouses rated as having high blame and a "low blame" LM on the text from spouses rated as having low blame. We trained the LMs on unigram word frequency counts for simplicity and to avoid more complex smoothing procedures and data sparsity issues. Both LMs were smoothed via interpolation with a λ-weighted background (BG) LM trained on out-of-domain text:

$$\tilde{P}(W_t|B) = (1 - \lambda)P(W_t|B) + \lambda P(W_t|\text{BG}),\ \ 0 < \lambda < 1 \tag{5}$$

Since estimating the probability of the most likely path through the ASR word lattice may not be robust, we incorporated the probabilities of the 100 most likely ("N-best") paths through the lattice for each speaker turn. We assumed in our implementation that the N-best hypotheses were independent; see Equation 6, where the n subscript refers to the n^{th} most likely path. In practice, we applied Equation 7 for numerical

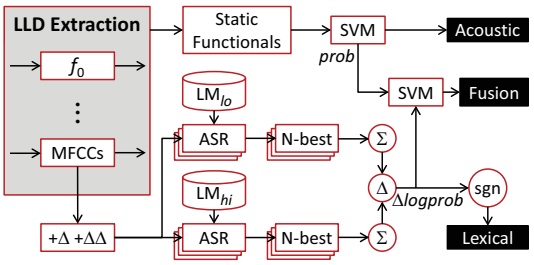

Figure 1: *System block diagram, from the low-level descriptors (LLDs) to the blame class outputs for the static acoustic classifier, ASR-derived lexical classifier, and fusion classifier; see Section 4 for details.*

reasons. See Figure 1 for a depiction of the ASR-derived lexical classifier, where we denote the smoothed LMs for low and high blame as LM_{lo} and LM_{hi}, respectively.

$$B^* = \underset{B,W}{\operatorname{argmax}} \prod_n \prod_t P(O_t|W_{t,n})\tilde{P}(W_{t,n}|B) \tag{6}$$

$$= \underset{B,W}{\operatorname{argmax}} \sum_n \sum_t \log P(O_t|W_{t,n})\tilde{P}(W_{t,n}|B) \tag{7}$$

4.3. Oracle Lexical Classifier

To find an upper bound on the performance of the proposed ASR-derived lexical classifier, we ran an oracle experiment that assumed we had perfect word recognition rate (i.e., we used the manual transcription). This oracle classifier is shown in Equation 8, where W is the sequence of transcribed words across the session for the rated spouse, and we used the same smoothed LMs as in Section 4.2 to compute $\tilde{P}(W|B)$.

$$B^* = \underset{B}{\operatorname{argmax}}\ \tilde{P}(W|B),\ \ B \in \{-1, 1\} \tag{8}$$

4.4. Fusion Classifier

Fusion of multimodal information has been advantageously applied in many engineering research domains. For example, improved emotion recognition has been reported when fusing audio/language/discourse features [1] and audio/video features [19]. Fusion typically takes place at the feature-level (e.g., by combining features at the input of a classifier), score-level (e.g., by combining output confidence scores from many classifiers), or decision-level (e.g., by voting on multiple classifier decisions). For our experiments, fusion at the score-level was most applicable, given the high dimensionality of the static acoustic classifier (not ideal for feature-level fusion) and since we only had two classifiers (not ideal for decision-level fusion).

The fusion features FF were computed using Equation 9, where conf_c is a non-negative confidence score for classifier c:

$$FF_c = (\text{conf}_c)(B_c^*),\ \ B^* \in \{-1, 1\},\ \text{conf} \geq 0 \tag{9}$$

For the ASR-derived and oracle lexical classifiers, the magnitude of the difference in log-probabilities between the competing LMs served as the confidence score. For the static acoustic SVM classifier, class probability estimates (made by LIBSVM using internal cross-validation on the training data) were the confidence scores [17].

We again used LIBSVM's SVM for the fusion classifier and z-normalized the fusion features, so they were on a comparable scale. We tried three pairs of classifier combinations: fusing the static acoustic and ASR-derived lexical classifiers (see Figure 1), fusing the static acoustic and oracle lexical classifiers, and fusing the two lexical classifiers.

5. Results & Discussion

Table 2 shows the performance of the various classifiers on the 280 instances. Using a difference in binomial proportions statistical test, we see that all proposed classifiers had significantly higher accuracy than chance accuracy of 50% (all $p < 0.01$). All oracle classifiers had significantly higher accuracy than all non-oracle classifiers (all $p < 0.01$), with no statistical difference between any of the oracle classifiers (all $p > 0.05$). There was no statistically significant difference between any of the non-oracle classifiers ($p > 0.05$), except the acoustic and ASR-derived lexical fusion classifier had significantly higher accuracy than the ASR-derived lexical classifier alone ($p < 0.05$).

In isolation, the oracle lexical classifier (which uses the perfect transcription) performed best, which suggests that lexical information is critical for classifying blame behaviors; this agrees with both intuition and the coding manual [8]. Even though the static acoustic classifier ignores these important lexical cues, it outperformed the ASR-derived lexical classifier, although not significantly ($p > 0.05$). Achieving 75% classification accuracy with the ASR-derived lexical classifier is a promising result, especially considering the noisy acoustic conditions and spontaneous nature of the corpus.

The significant difference between the ASR-derived and oracle lexical classifiers can most likely be attributed to the quality of the ASR word lattices. We found the ASR word error rate ranged from 40%-90% across the sessions (using standard metrics on the most likely word hypothesis). For less noisy data, we would expect the quality of the ASR lattices to improve and the classification performance to increase.

We see from the fusion experiments that performance decreased when we fused the two lexical classifiers, most likely because both of these classifiers model the language use of the spouses. We got a 0.7% absolute (0.8% relative) improvement when we fused the static acoustic classifier with the oracle lexical classifier. Although this difference is not significant, it suggests that the system was able to incorporate complementary acoustic information from the spouses' speech.

Although it is not a statistically significant difference in performance ($p > 0.05$), we saw a 2.5% absolute (3.1% relative) boost in performance when we fused the static acoustic and ASR-derived lexical classifiers. This fusion classifier advantageously combined automatically derived blaming cues from the spouses' speech and language. It also has the benefit of incorporating confidence scores, which can be interpreted to determine the relative importance of "what the spouse said" versus "how the spouse spoke," with respect to the perception of blame.

6. Conclusions & Future Work

Using a corpus of married couples' interactions, we showed we could successfully separate 82% of the extreme instances of blaming behavior conveyed by the spouses through fusion of automatically derived speech and language information. In the future, we will work to improve: the static acoustic classifier (e.g., by implementing feature selection techniques); ASR-based lexical classifier (e.g., by training "blame class"-dependent acoustic models and experimenting with other procedures to merge N-best hypotheses); and fusion classifier (e.g., by experimenting with new confidence score estimation schemes).

Blaming behaviors are an important cue to detect because of their significance in the context of couple therapy. Automatically detecting blaming behaviors in marital discussions could facilitate clinician-guided drill-down therapy sessions to reduce the occurrence of this negative behavior. As part of our future work, we also plan to extend these fusion experiments to other behavioral codes. In addition, we will apply these behavioral signal processing methodologies to other health-care related domains, such as autism and addiction.

System	*Classifier*	*Acc (%)*
Baseline	Chance	50.0
Unimodal	Acoustic	**79.6**
	Lexical/ASR	**75.4**
	Lexical/Oracle	91.1
Fusion	Acoustic + Lexical/ASR	**82.1**
	Acoustic + Lexical/Oracle	91.8
	Lexical/ASR + Lexical/Oracle	87.5

Table 2: *The accuracy of the proposed classification methods.*

7. Acknowledgements

This research was supported in part by the National Science Foundation and the Viterbi Research Innovation Fund. Special thanks to the Couple Therapy research staff for sharing the data.

8. References

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

[2] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[3] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Proc. Interspeech*, 2009.

[4] D. Jurafsky, R. Ranganath, and D. McFarland, "Extracting social meaning: Identifying interactional style in spoken conversation," in *Proc. Human Language Technologies*, 2009.

[5] B. Weiss and F. Burkhardt, "Voice attributes affecting likability perception," in *Proc. Interspeech*, 2010.

[6] M. P. Black, A. Katsamanis, C.-C. Lee, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *Proc. Interspeech*, 2010.

[7] C.-C. Lee, M. P. Black, A. Katsamanis, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proc. Interspeech*, 2010.

[8] C. Heavey, D. Gill, and A. Christensen, *Couples interaction rating system 2 (CIRS2)*, University of California, Los Angeles, 2002. [Online]. Available: http://christensenresearch.psych.ucla.edu/

[9] S. Dimidjian, C. R. Martell, and A. Christensen, *Clinical Handbook of Couple Therapy*, 4th ed. The Guilford Press, 2008, ch. Integrative behavioral couple therapy, pp. 73–106.

[10] A. Christensen, D. Atkins, S. Berns, J. Wheeler, D. H. Baucom, and L. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *J. of Consulting and Clinical Psychology*, vol. 72, pp. 176–191, 2004.

[11] P. K. Ghosh, A. Tsiartas, and S. S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[12] A. Katsamanis, M. P. Black, P. G. Georgiou, L. Goldstein, and S. S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Very-Large-Scale Phonetics Workshop*, Jan. 2011.

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

[14] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE - The Munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, 2010, pp. 1459–1462.

[15] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The Interspeech 2010 paralinguistic challenge," in *Proc. Interspeech*, 2010.

[16] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?" in *Proc. ICASSP*, 2008.

[17] C. C. Chang and C. J. Lin, *LIBSVM: A library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[18] E. Ettelaie, P. G. Georgiou, and S. S. Narayanan, "Cross-lingual dialog model for speech to speech translation," in *Proc. Interspeech*, 2006.

[19] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *Interspeech*, 2010.