Building A Naturalistic Emotional Speech Corpus by Retrieving Expressive Behaviors From Existing Speech Corpora

Soroosh Mariooryad, Reza Lotfian and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

soroosh.ooryad@utdallas.edu, reza.lotfian@utdallas.edu, busso@utdallas.edu

Abstract

A key element in affective computing is to have large corpora of genuine emotional samples collected during natural conversations. Recording natural interactions through telephone is an appealing approach to build emotional databases. However, collecting real conversational data with expressive reactions is a challenging task, especially if the recordings are to be shared with the community (e.g., privacy concerns). This study explores a novel approach consisting in retrieving emotional reactions from existing spontaneous speech databases collected for general speech processing problems. Although most of the recordings in these databases are expected to have non-emotional expressions, given the naturalness of the interactions, the flow of the conversation can lead to emotional responses from conversation partners which we aim to retrieve. We use the IEMOCAP and SEMAINE databases to build emotion detector systems. We use these classifiers to identify emotional behaviors from the FISHER database, which is a large conversational speech corpus recorded over the phone. Subjective evaluations over the retrieved samples demonstrate the potential of the proposed scheme to build naturalistic emotional speech database.

Index Terms: emotion recognition, expressive speech, information retrieval, emotional databases

1. Introduction

Collecting naturalistic emotional database for affect analysis is a challenging task, especially if the scope is to capture subtle emotions characterizing real-life behaviors. Ethical and legal issues constrain the available approaches to collect real emotion expressions. Hence, relying on actors reading sentences in specific emotion became one of the early approaches to record emotional speech corpora [1–3]. However, this method often results in exaggerated expressions, which can not characterize subtle behaviors [4]. Given the differences between spontaneous and read speech, this approach does not represent natural conversational speech. Simulating a conversation between two or more speakers can help to elicit more conversational-like speech. Variations of this technique include using scripts or hypothetical situations (IEMOCAP database [5]), performing a collaborative task (Recola database [6]) or eliciting emotions with the *sensitive artificial listener* (SAL) technique, in which a conversation partner plays the role of an agent with a given personality (SEMAINE database [7, 8]). All these methods increase the cost and complexity of the recording. Therefore, it became popular the use of data recorded in uncontrolled setting during natural conversations. Examples of these databases include conversational speech recorded in call centers [9, 10], interaction of kids with robots (FAU-AIBO database [11]), TV talk-shows (VAM database [12]) and interviews/video blogs [13–15].

This study explores a novel approach to collect naturalistic emotional databases. Given the large body of effort in collecting conversational speech for *automatic speech recognition* (ASR) and *speaker identification* (SID), we propose to use an speech emotion detection system to retrieve emotional behaviors from these databases. Although these corpora were expected to capture nonemotional speech, the flow of the conversation during natural interaction between speakers likely elicited emotional responses full of frustration, excitement, sadness and anger, among other emotions. Therefore, retrieving emotional behaviors using emotional speech detectors represents a feasible, attractive and cost-efficient solution.

We use the IEMOCAP [5] and SEMANE [8] databases as the two large emotion recognition corpora to build the emotion detector system using acoustic features. Then, we retrieve emotional behaviors from the Fisher English database Phase 1 which contains over 5000 phone conversation [16]. The Fisher corpus was recorded for speech recognition purposes. We emotionally evaluate the expressive behaviors retrieved from this corpus. The subjective evaluations show the benefits of the proposed method, which achieves high precision rates in emotion detection. This method can help building large naturalistic emotional speech corpora without expending resources in data recording

2. Databases

The study considers three databases. We use two of them to train emotional speech detectors, and the remaining one to retrieve expressive behaviors. This section describes the corpora.

2.1. The IEMOCAP database [5]

The *interactive emotional dyadic motion capture* (IEMOCAP) database [5] was collected at the *University of Southern California* (USC) to study expressive human behaviors. Ten trained actors were recorded in five dyadic sessions. The emotions were elicited with scripts and spontaneous improvisations to evoke sadness, happiness, anger and frustration. Other emotions were also elicited as dictated by the course of the conversation between the actors. Although the database is collected from actors, the elicitation techniques rooted in well-established theories and methods of theater provide emotional manifestations closer to natural iterations [17]. The corpus contains ap-

This work was funded by NSF (IIS-1217104, IIS-1329659) and Samsung Telecommunications America.

proximately twelve hours of data, which was manually segmented, transcribed and emotionally annotated with categorical and attribute-based labels at the turn level. The phonetic transcription is estimated with forced alignment (phone and word level). Each turn is annotated by three evaluators. This study considers turns in which three independent evaluators reached majority vote agreement. Turns with overlapped speech were excluded from the experiments. Since we aim to detect the emotional samples, the emotional categories are grouped into a single group resulting in 4375 emotional and 1121 neutral samples (5496 in total). Further information about the database is provided in Busso et al. [5].

2.2. The SEMAINE Database [8]

The study also uses the *sustained emotionally coloured machine-human interaction using nonverbal expression* (SEMAINE) database [8]. The corpus provides audiovisual recordings of natural interaction between a user and an operator using the *sensitive artificial listener* (SAL) framework [7]. In SAL, a person, or a virtual agent, interacts with the user using a predefined personality, aiming to elicit emotional reactions. This study only considers the interactions in which the operator was played by another human (solid SAL). This set includes 128 conversation sessions recorded from 18 unique speakers. Emotion evaluations are available for 83 of these sessions. The labels include time-continuous attribute evaluations using Feeltrace [18]. The number of evaluators varies across sessions from 2 to 8, who annotated arousal, valence, expectation and power, among other emotional descriptors. This study only uses arousal and valence which are the most prominent emotional attributes.

The sessions are segmented into speaking turns. We aggregate the results of the annotation by averaging the scores provided by the raters. Then, the average continuous ratings along the duration of each turn is assigned as a sentence level score (average across time and annotators). Therefore, each turn is represented by a pair of valence-arousal values between -1 and +1. To retrieve emotional behaviors, this study explores detectors, in which binary classifiers are trained to distinguish between neutral and emotional sentences. To set the binary labels to train the classifiers, we use the approach proposed in our previous studies [13, 19]. We define two circles centered at the origin of the valence-arousal coordinate system. The inner circle has radius equal to 0.3. All the samples included within this circle are considered as neutral (1215 samples). Notice that Feeltrace explicitly asks the raters to set the mouse cursor at the origin for neutral sentences [18]. The second circle has a radius equal to 0.4. All the turns with average scores lying outside this circle are considered as emotional turns (532 samples).

2.3. Fisher English Database [16]

The Fisher English Training Speech Part 1 Speech database is a conversational telephone speech corpus of American English [16]. This database was created by the *linguistic data consortium* (LDC) to develop robust *automatic speech recognition* (ASR) systems for conversational speech. We used this corpus to retrieve unsolicited emotional behaviors elicited during natural spontaneous interactions. During the Fisher data collection a robot randomly placed calls to two participants and asked them to conduct a conversation about an assigned topic. This study uses the phase 1 of the Fisher corpus, which includes over 5000 telephone conversations on 40 topics. Each conversation lasts up to 10 minutes. The conversations are segmented into speaking turns resulting in more than 800,000 segments. The large size of the corpus, and the protocol to collect spontaneous phone conversations make this corpus an ideal candidate to demonstrate the benefits of our method. Since this database was designed for ASR applications, it does not contain emotional annotations.

Table 1: Number of sentences retrieved under different settings. Due to overlapped sets, there are only 536 distinct sentences.

Set	# Turns	Corpus (Training)	Training Data
Emotional Sets	top 100	IEMOCAP	Balanced
	top 100	IEMOCAP	Unbalanced
	top 100	SEMAINE	Balanced
	top 100	SEMAINE	Unbalanced
	top 100	Fusion	Balanced
	top 100	Fusion	Unbalanced
Neutral Sets	top 50	IEMOCAP	Unbalanced
	top 50	SEMAINE	Unbalanced
Random Set	50	–	–

3. Retrieving Expressive Behaviors

We propose to build emotion detectors by training separate models using independent emotional corpora (IEMOCAP and SEMAINE). The final decision is achieved by fusing the scores derived from both models. The classifiers are build with linear kernel *support vector machine* (SVM) trained with *sequential minimal optimization* (SMO). We use the implementation provided by the WEKA toolkit [20]. The SVM complexity parameter is set to $c = 0.1$ for all settings.

This study uses the common feature set provided for the speaker state challenge at INTERSPEECH 2011 [21]. This set includes 4368 *high level descriptors* (HLDs) extracted using the OpenSMILE toolkit [22]. The set includes an exhaustive number of prosodic, spectral and voice quality features. We use a two-level feature selection approach to reduce the dimension of the feature vector. Since implementing feature selection that maximizes the performance of classifiers is computationally expensive for this large feature set, first, we use forward feature selection based on inter-class and intra-class distance measure to reduce the number of features to 500. This method iteratively extends the current feature set by adding the feature which best minimizes the intra-class distances while maximizing the inter-class distances. Then, we select 100 features by maximizing the performance of the SVM classifier using *forward feature selection* (FFS). For consistency, the feature selection is performed only on the SEMAINE database. Then, the 100 features are used to build emotion detection models using the SEMAINE and IEMOCAP databases. For each of the three corpora the acoustic features are normalized per speaker using z-normalization to compensate for speaker variability.

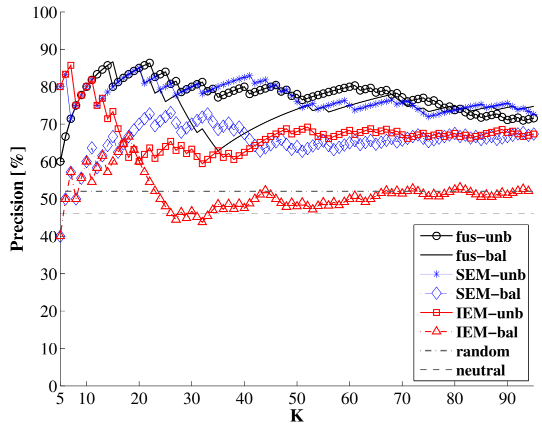
As described in Section 2, both expressive databases are emotionally unbalanced. The IEMOCAP database has more emotional samples and the SEMAINE database has more neutral samples. We expect that most of the sentences in the Fisher database are emotionally neutral. We evaluate the effect of balancing the training data on the emotional retrieval system. We use undersampling to balance the training sets. We evaluate the turns in the Fisher database using the emotion detectors built with balanced and unbalanced training set from either IEMOCAP or SEMAINE corpora. We only consider sentences with duration 5s or longer (157,959 samples). The confidence of the classifiers is used to select the top 100 emotional samples from the Fisher database under each of these settings. We also select the top 50 neutral samples from the classifiers built with unbalanced training sets. Likewise, we investigate the fusion of the classifiers trained with these two databases by averaging their likelihoods. We select the top 100 emotional sentences after fusion. Finally, a set of 50 utterances are randomly selected as a reference. Table 1 summarizes the different settings and the corresponding number of turns selected to create the corpus. The different settings provide 600 emotional, 100 neutral and 50 random samples, resulting in 750 utterances in total. However, due to the overlap between these sets the actual number is only 536.

4. Analysis of Retrieved Emotional Content

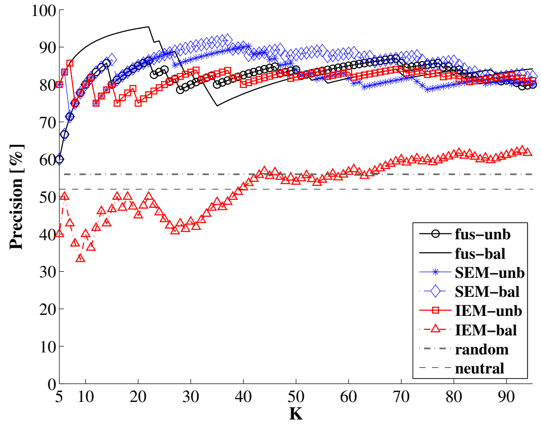
We use *Amazon mechanical Turk* (AMT) to evaluate the emotional content of the retrieved audio samples. In total, 90 different turkers evaluated the set of 536 utterances providing five evaluations per sentence. In preliminary studies using crowdsourcing, we noticed that the inter-evaluator agreement in annotating emotional labels increases when the turkers evaluate more than one video per *human intelligence task* (HIT). The turkers can calibrate their assessment by evaluating multiple videos per task. Therefore, we include 30 sentences per HIT. After listening to a sentence, the turkers complete a questionnaire with three parts. The first part consists of a five-point Likert scale capturing the degree of emotion in the sentence (neutral versus emotional). The second part consists of categorical emotional labels (angry, happy, neutral, sad, frustrated, surprised, fearful, depressed, excited, disgusted and other). The turkers were instructed to select all the emotional classes that represent the emotional content in the audio. The third part evaluates the emotional content in terms of arousal, valence and dominance. We use the *self-assessment manikins* (SAMs) [23,24] to visually guide the evaluators in annotating these dimensional attributes.

Since the Fisher database does not have emotional labels, we rely on the precision rate estimated from the retrieved emotional behaviors (e.g., correct samples retrieved by the system). For this purpose, we use the *precision at K* (P@K) metric. We sort the retrieved sentences in descending order based on their likelihood of being *emotional*. Then, we measure the precision by considering only the first K retrieved samples (i.e., K most emotional samples). Figure 1 reports the results for $K \in [5, \ldots, 100]$. The ground truth labels are set based on the subjective evaluations. We use two criteria to assign the labels. The first method applies a threshold over the average emotional scores assigned by the evaluator (i.e., five-point Likert scale). The ratings are mapped into numeric values between 1 (neutral) and 5 (emotional). We consider a sentence as *emotional* if its average score across turkers is above 2.5. Otherwise, the sentence is considered as *neutral*. Notice that a score equal to 3 implies that evaluators perceived emotional traits in the sentence, so we set a threshold lower than 3. Figure 1(a) gives the results under this criterion. The second approach to assign the labels is based on the percentage of neutral labels assigned by the turkers to each sentence in the categorical evaluation. The labels of the samples are set to *emotional* if this percentage is less than 50%. Figure 1(b) gives the results under this criterion. As a reference, Figures 1(a) and 1(b) also show the fraction of emotional samples, according to the subjective evaluations, from the random and neutral sets under these thresholds (constant dashed lines).

The figures show high precision rates for both criteria. The precision rates for different settings are similar, and significantly higher than the ones for neutral and random sets. The



(a) Labels are set with scale-based emotion evaluations.

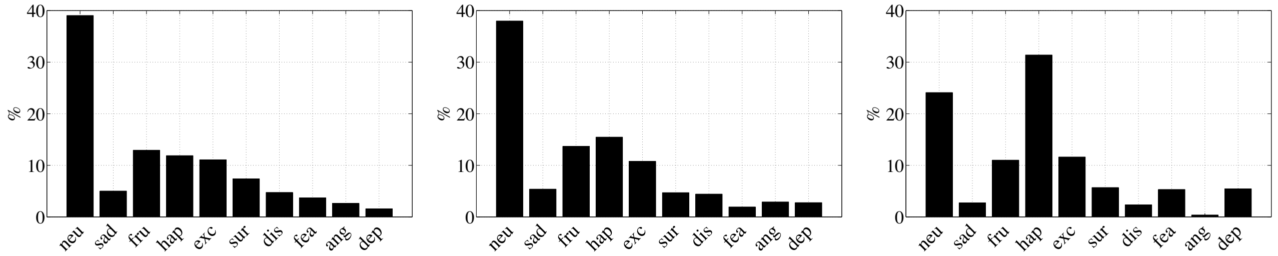


(b) Labels are set with categorical label evaluations

Figure 1: Precision of retrieving the K most emotional samples. The neutral/emotional labels are set with (a) scale-based emotion evaluations, and (b) categorical labels evaluations.

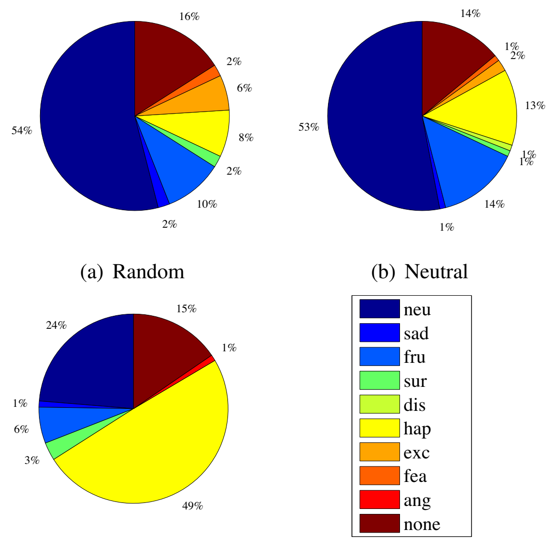
models built with balanced training sets achieve lower precision rates especially for the IEMOCAP database. Notice that the feature selection is only performed on the SEMAINE database, which can explain the lower performance. Overall, the fusion of the models trained with unbalanced partitions gives the best precisions. Notice that the neutral set has fewer emotional samples than the random set, which indicates that the emotion detection models are not only capable of detecting the emotional samples, but also can identify neutral instances. The rest of the evaluations consider the retrieved sentences using the fusion approach, which are compared with the sentences from the neutral and random sets.

Figure 2 depicts the histogram of the categorical labels assigned to the three sets before defining consensus labels for the sentences. The first group includes 50 randomly selected samples (Figure 2(a)). The second group includes the top 100 neutral samples selected by our classifiers (Figure 2(b)). The third group consists of the top 100 emotional samples retrieved by the fusion of the emotion detectors trained with unbalanced datasets (Figure 2(c)). The figure shows that the neutral and random sets exhibit similar distributions. This result shows that most of the samples in the Fisher corpus are emotionally neutral, which is expected. However, we did not expect to observe that more than half of the random sentences convey emotional traits. This result shows that building an emotional corpus from existing



(a) Random (b) Neutral (c) Emotional (fusion/unbalanced)

Figure 2: Histogram of categorical emotional labels assigned to (a) 50 random sentence, (b) top 100 neutral samples identified by the IEMOCAP and SEMAINE emotion detectors (50 each), and (c) top 100 emotional samples identified by fusing the emotion detectors.



(a) Random (b) Neutral

(c) Emotional

Figure 3: Consensus labels assigned using majority vote. The emotional set has the top 100 emotional sentences retrieved by fusing the detectors, trained with unbalanced partitions.

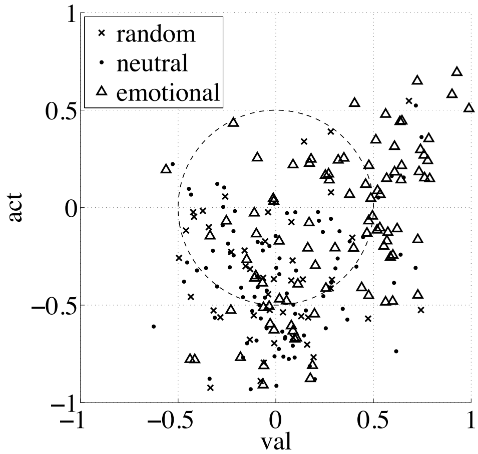


Figure 4: Distribution of the retrieved samples in the valence-arousal space. The results are reported for neutral, random and emotional (fusion/unbalenced) sets.

spontaneous conversational corpora is an appealing approach. The emotional samples retrieved by the emotion detectors follow a completely different distribution. The most frequently selected emotion is happiness, which reflects the colloquial protocol used to record the Fisher corpus. We noticed that many of these sentences contain laughter. We also observe depressive, frustrated and excited behaviors.

Figure 3 shows the distribution of the annotated emotions after defining the consensus label per sentence using majority vote. The category *none* corresponds to sentences without agreement between turkers. This figure also clearly shows the effectiveness of the proposed method in identifying the emotional samples. Finally, Figure 4 shows the scatter plot of the samples in the valence-arousal space after estimating the average scores assigned to each sentence. Given the protocol to collect the Fisher database, many of the samples have positive valence. This figure shows that the emotional classifiers are effective in detecting samples with high arousal and high valence.

5. Discussion and Conclusions

This study introduced a new approach to build naturalistic emotional speech corpora by retrieving emotional behaviors from existing spontaneous conversational databases recorded for various speech processing problems (e.g., ASR, SID). Perceptual evaluations over turns selected at random revealed that approximately 30% of the sentences received emotional labels after majority vote. The large percentage of sentences conveying emotions in the Fisher corpus validates our approach of retrieving emotional sentences from existing conversational corpora. The emotion detection models trained with available emotional corpora are effective to retrieve these expressive sentences.

This study opens new opportunities to create big emotional databases with natural, spontaneous expressive behaviors. An important direction is to retrieve emotionally balanced behaviors. The retrieved samples conveyed mostly positive emotions (happiness, excitement), since the current retrieval system does not distinguish between emotions (neutral versus emotional problem). We are building specialized emotion detectors to retrieve specific emotions. We are also buildings detectors to retrieve sample with specific arousal-valence values.

Fusing emotion detectors was effective in retrieving expressive behaviors. We can extend the approach by building multiple retrieval systems that are expert on specific emotional dimensions by using various emotional databases and acoustic and prosodic features. This approach provides multiple views of the data and enables the use of co-adaptation techniques [25]. Since we have a set of data for each speaker, unsupervised normalization techniques can also be employed to better compensate for speaker and recording variabilities. In particular, we are planning to explore the use the *iterative feature normalization* (IFN) [14, 26] scheme, which has shown good performance in paralinguistic recognition tasks [13, 27].

6. References

[1] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts," Philadelphia, PA, USA, 2002, Linguistic Data Consortium.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 1517–1520.

[3] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Sixth International Conference on Multimodal Interfaces ICMI 2004*. State College, PA: ACM Press, October 2004, pp. 205–211.

[4] E. Douglas-Cowie, L. Devillers, J. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox, "Multimodal databases of everyday emotion: Facing up to complexity," in *9th European Conference on Speech Communication and Technology (Interspeech'2005)*, Lisbon, Portugal, September 2005, pp. 813–816.

[5] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[6] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013.

[7] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, "The sensitive artificial listener: an induction technique for generating emotionally coloured conversation," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 1–8.

[8] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

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

[10] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Interspeech - International Conference on Spoken Language (ICSLP)*, Pittsburgh, PA, USA, September 2006, pp. 801–804.

[11] S. Steidl, "Automatic classification of emotion-related user states in spontaneous childrens speech," Ph.D. dissertation, Universität Erlangen-Nürnberg, Erlangen, Germany, January 2009.

[12] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 865–868.

[13] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 5117–5120.

[14] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 386–397, October-December 2013.

[15] L. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces (ICMI 2011)*, Alicante, Spain, November 2011, pp. 169–176.

[16] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *International conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, May 2004.

[17] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[18] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, Northern Ireland, UK: ISCA, September 2000, pp. 19–24.

[19] J. Arias, C. Busso, and N. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Computer Speech and Language*, vol. 28, no. 1, pp. 278–294, January 2014.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, June 2009.

[21] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, August 2011, pp. 3201–3204.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[23] L. Fischer, D. Brauns, and F. Belschak, *Zur Messung von Emotionen in der angewandten Forschung*. Pabst Science Publishers, Lengerich, 2002.

[24] 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, October-November 2007.

[25] G. Tur, "Co-adaptation: Adaptive co-training for semi-supervised learning," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, Taipei, Taiwan, April 2009, pp. 3721–3724.

[26] C. Busso, A. Metallinou, and S. Narayanan, "Iterative feature normalization for emotional speech detection," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 5692–5695.

[27] D. Bone, M. P. Black, M. Li, A. Metallinou, S. Lee, and S. Narayanan, "Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors," in *12th Annual Conference of the International Speech Communication Association (Interspeech'2011)*, Florence, Italy, August 2011, pp. 3217–3220.