Abstract
In this paper we present an innovative approach for utterance-level emotion recognition by fusing acoustic features with lexical features extracted from automatic speech recognition (ASR) output. The acoustic features are generated by combining: (1) a novel set of features that are derived from segmental Mel Frequency Cepstral Coefficients (MFCC) scored against emotion-dependent Gaussian mixture models, and (2) statistical functionals of low-level feature descriptors such as intensity, fundamental frequency, jitter, shimmer, etc. These acoustic features are fused with two types of lexical features extracted from the ASR output: (1) presence/absence of word stems, and (2) bag-of-words sentiment categories. The combined feature set is used to train support vector machines (SVM) for emotion classification. We demonstrate the efficacy of our approach by performing four-way emotion recognition on the University of Southern California’s Interactive Emotional Motion Capture (USC-IEMOCAP) corpus. Our experiments show that the fusion of acoustic and lexical features delivers an emotion recognition accuracy of 65.7%, outperforming the previously reported best results on this challenging dataset.

Index Terms: emotion recognition, model-based acoustic features, lexical features

1. Introduction
Automatic assessment of the emotional state of an individual has a very important role in several applications ranging from affective computing for virtual training systems [1] to early diagnosis of psychological health disorders [2]. Studies have shown that information relevant for interpretation of speaker’s emotional state is contained both in linguistic content and acoustic paralinguistic properties of speech. In some cases, linguistic content may not be emotionally rich (i.e. it is very difficult to recognize emotion from the transcript), and we rely on subtle speech characteristics: pitch, loudness, voicing patterns, etc. for assessing a speaker’s emotional state. In other cases, speech is acoustically neutral and the meaning and associated sentiments inferred from the linguistic content are the main cues for the emotion recognition. Therefore, humans actively make decisions about importance of both “What is said?” and “How is it said?”.

Accuracy of automatic emotion recognition from speech depends largely on the choice of informative and meaningful features. Acoustic features play a dominant role in the emotion recognition literature. Gaussian mixture models that include segmental (energy, mel-frequency cepstral coefficients (MFCC), formants) and supra-segmental (pitch and degree of voicing) frame-level descriptors [3]. It is an accepted practice to generate fixed-dimensional vectors for utterance-level emotion classification by computing various statistical functionals (mean, standard deviation, range etc.) over a variable number of frame-level descriptors extracted from an utterance. On the other hand, linguistic features used for emotion recognition include presence indicators of lexemes [4], n-grams [5], and various bag-of-words representations [6]. Since automatic speech recognition (ASR) of emotional speech is a difficult problem, most of the work is based on reference transcripts [4, 6], while a small number of studies rely on ASR [7, 8].

In this paper we present four-way utterance-level emotion (angry, happy, sad, neutral) recognition results on USC-IEMOCAP database [9]. Our main contributions are: (a) introduction of a feature class based on scoring of frame-level MFCCs by emotion dependent models (model-based features), and (b) analysis of emotion recognition performance based on feature-level fusion of acoustic and linguistic (lexeme and sentiment) features. We based our recognition experiments on three feature classes. The first class contains a set of basic frame-level features (energy, pitch, formants), voice quality features (jitter, shimmer), voicing statistics, and MFCCs. For features in the first category we calculated different statistical functionals of the frame-level features on the utterance-level. While this approach is appealing for slowly varying features, moments of the feature value distributions could oversimplify utterance-level representations of highly non-stationary features (e.g. MFCCs). To overcome this drawback, we propose a model-based feature set obtained by scoring all MFCCs within an utterance by emotion-dependent Gaussian mixture models (GMM). We further normalize score vectors and fuse them in three ways by: (a) calculating mean of the normalized score vector on the utterance-level, (b) generating histograms by voting for the highest scoring emotion model, and (c) estimating parameters of the Dirich-
let distribution that generates normalized scores within an utterance. Related work include GMM supervector approaches [10] and our previous work on parametric MFCC features [11]. We augment the described acoustic features by two types of linguistic features derived from the ASR output: binary word-stem presence indicators and sentiment features based on bag-of-words sentiment categories. We conducted all experiments in a leave-one-speaker-out configuration, a standard approach for the USC-IEMOCAP dataset. For emotion recognition from acoustic features we get the best accuracy reported on the dataset (60.2%) and further introduction of ASR based lexical features boosts recognition accuracy to 65.7%.

The remainder of the paper is organized in the following way. Section 2 provides a detailed description of acoustic, model-based and lexical features used in our system. In Section 3, we briefly describe the USC-IEMOCAP dataset. Section 4 details our experimental methodology and summarizes key results. Section 5 concludes the paper and provides directions for future research in this area.

2. Features

To obtain cues important for emotion recognition both from acoustic and lexical content we extracted three different categories of features. The first class includes acoustic low level descriptors (energy, pitch, first two formants, and ASR output). We removed stop-words and applied Porter stemming to remove common morphological and inflectional endings. Additionally, we augmented the INTERSPEECH 2012

2.1. Low-level Acoustic Descriptors

We use numerous utterance-level acoustic features derived as statistical functionals (maximum, mean, standard deviation, range) of frame-level features: F0, intensity, first two formants and 36-dimensional MFCCs with delta and acceleration values. These features are well-documented and widely used for emotion recognition [3]. In order to mitigate effects of variability in speaker characteristics and recording conditions, we normalize the energy mean of speech signal, and perform cepstral mean normalization of MFCCs for each speaker in the dataset. Additionally, we extract features related to two voice quality measures, jitter and shimmer, that are correlated with negative emotions [2]. Jitter is defined as a relative difference between consecutive periods in a voiced speech segment. Shimmer measures the degree of local change in the speech intensity. Both low-level acoustic and voice quality features were extracted on 25ms processing frames with 10ms frame shift. Finally, we extract voicing statistics at the utterance level as intuitive indicators of excitement. These features include: (1) fraction of unvoiced frames in an utterance, (2) number of voice breaks (inter-pulse intervals longer than a certain threshold), and (3) the ratio of duration of all voice break intervals and the total duration of speech. All features were extracted using Praat [12], an open-source program for speech analysis. We have successfully used this feature set for emotion recognition on the Berlin Emotional Speech Corpus [11].

2.2. Model Distance Features

Our motivation is to transform segmental MFCC feature vector $x$ in a way that retains information relevant for emotion classification and avoid simplification through the direct computation of utterance level statistics.

Let us assume that $M$ models are trained on the original frame-level feature vector space. Choice of the models is open as long as they capture properties important for the classification task and can be used to score frame-level feature vectors. Let an utterance $x$ be represented by $T$ segmental feature vectors $x = (x_1, \ldots, x_T)$. We denote the likelihood that the feature vector $x_i$ is generated by a model $m$ as $l_i,m$, and denote the vector of model likelihoods as $l = (l_i,m : m = 1, \ldots, M)$. Normalization of the model likelihoods creates a vector of normalized model likelihoods $p_i = (p_i,m : m = 1, \ldots, M)$, where $p_i,m = \frac{l_i,m}{\sum_{m'} l_i,m'}$. Vectors of normalized likelihoods $p_i,i,m = 1, \ldots, T$ are multinomial distributions and we propose three ways to fuse them into a single fixed-dimensional utterance-level feature vector. First, we calculate distribution mean $p = \frac{1}{T} \sum_{i=1}^{T} p_i$. Second, we calculate histogram of votes for the highest scoring model $h = [h_1, \ldots, h_M]$, where:

$$h_m = \frac{1}{T} \sum_{i=1}^{T} \mathbb{I}(m = \text{argmax}_k p_i,k, k = 1, \ldots, M).$$

Finally, under an assumption that the multinomial distributions $p_i,i,m = 1, \ldots, T$ are i.i.d. samples from a Dirichlet distribution with parameters $\alpha = (\alpha_1, \ldots, \alpha_M)$ (Equation 1) it is possible to estimate parameters of the Dirichlet distribution in the maximum likelihood sense [13]. For this paper, instead of the maximum likelihood parameter estimates we have used their moment matching approximations [14].

$$p(p_i | \alpha_1, \ldots, \alpha_M) = \frac{\prod_{i=1}^{M} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{M} \alpha_i)} \prod_{i=1}^{M} \alpha_{n_i,m,i}^{\alpha_{i,n_i,m,i}-1}$$

On the emotion recognition task we use two sets of $(p, h, \alpha)$ features. We fitted a $C$-component GMM for each emotion class of interest (angry, happy, sad, neutral) and obtained the first set of features by treating each GMM as one model. In this case un-normalized model likelihoods are:

$$l_{i,m,c} = \sum_{c=1}^{C} w_{m,c} p_i(x_i|m,c) = \sum_{c=1}^{C} w_{m,c} f_{m,c}(x_i),$$

where $w_{m,c}$ denotes component weights and $l_{i,m,c}(x_i)$ denotes class-component likelihoods. For the second feature set we assume that each component of emotion category GMMs is a model. Indexing these models by ordered pairs of emotion class - mixture component $(c, c)$, un-normalized model likelihoods are $w_{m,c} p_i(x_i|c,c)$. Feature sets obtained this way are 4 dimensional ($E = 4$, first set) and 20 dimensional ($E = 4$, $C = 5$, second set).

2.3. Lexical Features

In order to capture content-based cues for emotion recognition, we extracted two types of lexical features, sentiment and word stem features, from reference transcripts and ASR output. We removed stop-words and applied Porter stemming to remove common morphological and inflectional endings. Additionally, we augmented the
set of stemmed words with non-speech tokens, such as laughter and sighs, present in reference transcripts and ASR output. We defined word stem features as binary indicators of presence or absence of all unigram tokens from the dataset in the processed utterance.

For each utterance, we extracted 125 features representing numbers of utterance words that belong to different sentiment bearing categories. Each category is represented by a bag-of-word stems based on Linguistic Inquiry and Word Count (LIWC) [15] and General Inquirer (GI) system [16] lexicons. The LIWC-derived categories include: positive emotions, optimism, negative emotions, anxiety and fear, anger, sadness, swear words; and the GI based categories include arousal, activation and valence. The LIWC and GI categories contain 2318 and 7293 word stems, respectively.

3. Dataset

In this paper we use the USC-IEMOCAP database [9] to demonstrate the efficacy of our approach. This database contains 12 hours of data, consisting from multiple 5min dyadic interaction between professional female-male actors. Each interaction represents either acting out scripted play or engaging in spontaneous dialog. The dataset provides audio, video and motion capture trajectories of facial markers. Each interaction is manually split into utterances roughly corresponding to speaker turns in the dialogue, manually transcribed and annotated by human annotators. Utterances are labeled by at least three annotators with categorical emotion labels (anger, happiness, etc.) and emotion attributes (valence, activation and dominance on the scale 1-5).

In order to match experimental conditions in previously reported categorical emotion recognition studies on USC-IEMOCAP [17, 18], we consider all utterances on which labelers achieved majority vote agreement on one of five categories: anger, happiness, sadness, excitement and neutral. The happiness and excitement categories were further merged under the happiness label, forming the four categories used in our emotion recognition experiments. Each emotional category contains approximately equal number of sample utterances in the set of 5531 utterances used for recognition.

4. Experimental Results

We conducted all evaluations in a leave-one-speaker-out 10-fold cross-validation configuration. For each cross-validation split we extracted acoustic features for both test and train utterance sets. We modeled MFCCs of each emotion category with 5-mixture GMMs using only the training set feature vectors. Using these models we generated model-based MFCC features as described in Section 2.2 for all test and train utterances. Additionally, we decoded each utterance using ASR system with acoustic models trained on 1700 hours of broadcast news data and 65K word dictionary. Due to a mismatch between the acoustic models and spontaneous speech style of USC-IEMOCAP (including speaker overlaps) the ASR word error rate we got was high (40%). But using the manual transcriptions and ASR output we extracted lexical features as described in Section 2.3. From training set words we formed a dictionary of word-stems and created a binary vectors of indicators of their presence in test and train utterances. Also, we generated 125-dimensional vectors with coordinates representing counts of words that belong to each of 125 sentiment categories described by bags-of-words.

In the remainder of the paper we denote the utterance-level acoustic features (excluding MFCCs) as ACO, the proposed model-based acoustic features as M-MFCC and the lexical features derived from reference transcripts and ASR, respectively, as LEX-T and LEX-ASR.

We concatenated vectors of all features used in a particular experiment and formed test and train feature sets. We further split each training set into 9 single speaker folds to optimize parameters of the radial basis function kernel support vector machine (RBF-SVM) classifier. To maximize cross-validation classification accuracy on 9 training set folds we performed optimization over two parameters, parameter $C \in \{2^{-1}, \ldots, 2^8\}$ that controls trade-off between training errors and SVM margin maximization and the kernel width parameter $\gamma \in \{2^{-2}, \ldots, 2^{-2}\}$. Optimal classifier parameters were used to train RBF-SVM classifier on the full training set.

In Table 1 we present unweighted and weighted emotion classification accuracies for different feature sets. The unweighted accuracy is defined as mean accuracy over different emotion categories, while the weighted accuracy is a weighted mean with weights proportional to the number of utterances in a particular emotion category. The combination of ACO and model-based MFCC (M-MFCC) features outperformed combination of ACO and MFCC features, providing the best reported USC-IEMOCAP four-emotion recognition accuracy (60.2%) based on acoustic features. Lexical features provided additional improvement to 65.7% when derived from the ASR output and 70.1% when based on reference transcripts.

Table 2 contains accuracies for different classes and provides insight in compatibility of acoustic and lexical features. Experiments with transcript-based lexical features show that introduction of the lexical features brings the best improvement in happiness and neutral categories. As previously mentioned, the happiness category is obtained by merging original happiness and excitement categories. While this merge balances category weights, the drawback is that in the acoustic feature space, excitement sub-category is much harder to distinguish from other categories than the happiness sub-category [9]. The transcript-based lexical features outperform acoustic features on happiness and neutral categories, and both transcript and ASR-based lexical features equalize accuracies on different classes when combined with acoustic features.

As mentioned in Section 3, utterances in the USC-
Table 2: Emotion category recognition accuracies for different feature sets

<table>
<thead>
<tr>
<th>Feature set</th>
<th>ang</th>
<th>hap</th>
<th>sad</th>
<th>neu</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACO</td>
<td>94.0</td>
<td>46.9</td>
<td>62.7</td>
<td>64.6</td>
</tr>
<tr>
<td>ACO+MFCC</td>
<td>61.8</td>
<td>49.1</td>
<td>66.3</td>
<td>58.4</td>
</tr>
<tr>
<td>ACO+M-MFCC</td>
<td>65.2</td>
<td>55.7</td>
<td>64.9</td>
<td>55.1</td>
</tr>
<tr>
<td>LEX-ASR</td>
<td>43.8</td>
<td>41.9</td>
<td>28.3</td>
<td>61.4</td>
</tr>
<tr>
<td>LEX-T</td>
<td>68.7</td>
<td>68.5</td>
<td>44.0</td>
<td>69.4</td>
</tr>
<tr>
<td>ACO+M-MFCC+LEX-ASR</td>
<td>70.0</td>
<td>63.8</td>
<td>69.2</td>
<td>59.8</td>
</tr>
<tr>
<td>ACO+M-MFCC+LEX-T</td>
<td>73.8</td>
<td>70.1</td>
<td>72.0</td>
<td>68.4</td>
</tr>
</tbody>
</table>

IEMOCAP dataset originate from scripted and improvised interactions. To further examine performance of lexical features, we performed additional experiments using leave one speaker out cross validation in two ways: (a) train on utterances originating in both improvised and scripted interactions and test on utterances from improvised interactions, and (b) train and testing on utterances from improvised interactions. In both cases we avoided appearance of the utterances based on the same script in both test and train sets. Results for this set of experiments are presented in Table 3.

Table 3: Recognition accuracies for different feature sets and train-test setups. (1) - train (scripted and improvised), test (improvised); (2) - train (improvised), test (improvised). UW - unweighted accuracy, W - weighted accuracy

<table>
<thead>
<tr>
<th>Feature set</th>
<th>UW</th>
<th>W</th>
<th>ang</th>
<th>hap</th>
<th>sad</th>
<th>neu</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACO+M-MFCC (1)</td>
<td>63.0</td>
<td>62.4</td>
<td>60.6</td>
<td>64.4</td>
<td>70.4</td>
<td>56.7</td>
</tr>
<tr>
<td>LEX-ASR (1)</td>
<td>31.4</td>
<td>36.2</td>
<td>21.8</td>
<td>31.6</td>
<td>18.1</td>
<td>54.1</td>
</tr>
<tr>
<td>ACO+M-MFCC+LEX-ASR (1)</td>
<td>64.0</td>
<td>63.6</td>
<td>61.6</td>
<td>67.2</td>
<td>68.8</td>
<td>58.3</td>
</tr>
<tr>
<td>ACO+M-MFCC (2)</td>
<td>62.6</td>
<td>64.6</td>
<td>47.4</td>
<td>65.9</td>
<td>74.8</td>
<td>62.3</td>
</tr>
<tr>
<td>LEX-ASR (2)</td>
<td>31.2</td>
<td>38.8</td>
<td>9.3</td>
<td>46.2</td>
<td>16.9</td>
<td>52.2</td>
</tr>
<tr>
<td>ACO+M-MFCC+LEX-ASR (2)</td>
<td>63.7</td>
<td>66.6</td>
<td>45.7</td>
<td>71.3</td>
<td>74.0</td>
<td>64.0</td>
</tr>
</tbody>
</table>

While the overall recognition accuracies for lexical features based on ASR decreased, conclusions derived from the experiments on the full message set remained to hold: category dependent accuracies were still the highest for the happiness and neutral categories, combination of lexical and acoustic features outperformed acoustic features boosting accuracies obtained using acoustic features on happiness and neutral categories.

5. Conclusions and future work

In this paper we presented results on categorical emotion recognition on the USC-IEMOCAP database. The proposed model-based features based on class-conditional models in combination with acoustic features provided state-of-the-art emotion recognition performance on this dataset. Lexical features that include word stem and sentiment-based indicators exhibited the best performance on happiness and neutral categories that were the lowest performing categories using acoustic features alone. Combination of the acoustic, model-based, and ASR-derived lexical features balanced classification accuracy for different emotions and boosted the unweighted accuracy to 65.7%.

Future work will focus on using different models for generating model-based features, combining visual features extracted from the motion capture streams, and identifying frame subsets which are the most relevant for utterance-level emotion recognition.

6. Acknowledgements

We would like to thank Professor Shrikanth Narayanan, SAIL-USC, for providing us access to the USC-IEMOCAP database.

7. References