ANGRY EMOTION DETECTION FROM REAL-LIFE CONVERSATIONAL SPEECH
BY LEVERAGING CONTENT STRUCTURE

Wooil Kim and John H. L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas, USA

ABSTRACT
This study proposes an effective angry speech detection approach by leveraging content structure within the input speech. A classifier based on an “emotional” language model score is formulated and combined with acoustic feature based classifiers including TEO-based feature and conventional Mel frequency cepstral coefficients (MFCC). The proposed detection algorithm is evaluated on real-life conversational speech which was recorded between customers and call center operators over a telephone network. Analysis on the conversational speech corpus presents a distinctive property between neutral and angry speech in word distribution and frequently occurring words. An improvement of up to 6.23% in Equal Error Rate (EER) is obtained by combining the TEO-based and MFCC features, and emotional language model score based classifiers.

Index Terms— angry speech detection, content structure, emotional language model, TEO-based feature, classifier combination

1. INTRODUCTION
Reliable stress/emotion detection can be used to increase performance of speech/speaker recognition systems for a range of applications. Angry speech detection, our focus in this study, can be effectively employed by industry for call center systems to improve customer service. In this study, an effective angry speech detection scheme is proposed by leveraging content structure for real conversational speech. Recently, extensive research has been conducted in the speech area to improve performance of stress/emotion classification by employing lexical/language information [1]-[4]. In our study, a language model score is used to formulate an independent classifier for angry speech detection in order to utilize content structure of input speech. The classifier based on the language model score (i.e., a lexical feature) is combined with acoustic feature based classifiers including TEO-based feature [5] and conventional MFCC. The proposed detection method is evaluated on real conversational speech recorded between customers and call center operators over a telephone network in the United States. We will also present a distinctive property of neutral and angry speech based on content structure by showing a comparison of word distribution and frequently occurring words.

2. REAL CONVERSATIONAL SPEECH CORPUS
In order to evaluate the anger speech detection system on real-life environments, actual conversation audio recordings were used in this study, which were obtained between operators and customers over a telephone network. The audio signal was sampled at 6 kHz. From the obtained audio samples, segments of the customer speech were identified as “neutral” or “angry” for their emotional state. A total of 28 speakers were collected consisting of 15 female and 13 male speakers respectively. 136 segments for neutral speech and 124 segments for angry speech were processed, resulting in a total of 260 segments. Each segment has 3-6 sec duration formulating +25 minutes in total. It was found that the obtained speech samples include severe real-life background noise and channel distortion due to the telephone network, which make effective angry speech detection challenging.

To observe the content structure of the obtained real conversational speech corpus, each speech sample was manually transcribed. The vocabulary seen in the transcripts consists of 724 words. Fig. 1 shows a comparison of word distributions which occur in (a) neutral and (b) angry speech, showing a considerable difference in the distributions; plot (c) reflects the difference between (a) and (b). Table 1 shows the top 15 most frequently occurring words for neutral and angry speech among the words which are seen “commonly” for both neutral and angry speech of the real-life data. Table 2 shows a comparison of the top 15 most frequently occurring words which are observed “exclusively” in neutral or angry speech. From the listed words, it can be seen that the neutral speech sentences are likely to include digits (ONE, TWO, ..., NINE, O), alphabets (C, B, E, L, etc.), and other words (FIRST, JULY, AUGUST), which are expected to be
(a) Neutral Speech

(b) Angry Speech

(c) Difference

Fig. 1. Word distributions of (a) neutral and (b) angry speech utterances, and (c) their difference.

Table 1. Top 15 most frequently occurring words “commonly” observed.

<table>
<thead>
<tr>
<th>Neutral</th>
<th>Angry</th>
</tr>
</thead>
<tbody>
<tr>
<td>I, SEVEN, ONE, MY, TWO, THREE, FIVE, EIGHT, OKAY, IS, FOUR, NINE, HAVE, O, BECAUSE</td>
<td>I’M, NOT, ME, I’VE, THAT, THIS, TO, GONNA, NO, ON, BE, BEEN, HERE, GO, HOW</td>
</tr>
</tbody>
</table>

Table 2. Top 15 most frequently occurring words “exclusively” observed.

<table>
<thead>
<tr>
<th>Neutral</th>
<th>Angry</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIX, E, L, C, CALLING, THINK, HAD, R, B, COURT, DRIVE, FIRST, JULY, SIGNED, AUGUST</td>
<td>PEOPLE, CAN’T, CUSTOMER, EVEN, HUNG, PERSON, DIFFERENCE, FEE, HELPFUL, LONGER, TALKED, THEM, TIMES, YOU’RE, AGREE</td>
</tr>
</tbody>
</table>

used to provide the operator with specific information such as address, telephone number, credit card number, date, and others. On other hand, the angry speech utterances are more likely to contain negative words (NOT, NO, CAN’T, EVEN, HOW), words used for complaints (CUSTOMER, HUNG, TALKED, TIMES, HELPFUL, LONGER, FEE) and others (THAT, THIS, HERE). We believe that such distinguished properties of neutral and angry speech in content structure (i.e., word distribution, frequently used words, etc) should provide a key factor to improve performance for angry speech detection.

3. TEO-CB-AUTO-ENV: CRITICAL BAND BASED TEO AUTOCORRELATION ENVELOPE

As one of the acoustic features for angry speech detection, our previously proposed TEO-CB-Auto-Env feature extraction method [5] is employed in this study. The TEO profile obtained from the Gabor bandpass filter output is segmented on a short-term basis, followed by an autocorrelation operation. Once the auto-correlation response is found, the area under the auto-correlation envelope is obtained and normalized. A single area coefficient is found which corresponds to each frequency band. The resulting vector of area coefficients has been shown to be large for neutral speech (i.e., speech has high “regularity”) and low for speech that is produced with irregular excitation structure (e.g., for speech under stress)[5][6].

4. PROPOSED ANGRY SPEECH DETECTION EMPLOYING EMOTIONAL LANGUAGE MODEL SCORE

In our recent study on angry speech detection, we have proposed two types of combination methods; (i) feature combination and (ii) classifier combination [7]. In the feature combination approach, the conventional MFCC feature vector is appended to the TEO-CB-Auto-Env feature vector. In the classifier combination approach, the classifier based on the TEO feature and the second classifier using an MFCC feature are composed at the decision stage by combining the likelihood scores from both classifiers with a scale factor. The TEO-CB-Auto-Env was originally designed to represent non-linear characteristics of the voiced sound production (e.g., vowels), so the TEO-based angry speech detection performance could depend on the ability of effective vowel sound detection. Since, in our study, the entire speech duration is used without prior knowledge of phonetic information of the input speech, we believe that a combination of MFCC and TEO features could be more effective at increasing performance.

As a preliminary approach to leveraging content structure, a language model score is used in this study, which is commonly employed for speech recognition. Based on an initial language model with a large vocabulary, we generate the “emotional” language models using the transcripts of neutral and angry speech respectively. Likelihood scores for neutral and angry language models are calculated for the input speech, and formulate a 2-dimensional feature vector for a “lexical” feature. In a similar manner as the acoustic feature based methods, a Gaussian Mixture Model (GMM) is employed for the classifier using the emotional language model scores as the input feature. The formulated classifier based on language model score is fused with the acoustic feature based classifier at the decision stage by combining the resulting likelihood scores. Fig. 2 illustrates the proposed angry
speech detection scheme by combining the Emotional Language Model Score (ELMS) based classifier.

5. EXPERIMENTAL RESULTS

5.1. Evaluation of Acoustic Features on Real-Life Conversational Speech

For performance evaluation of angry speech detection on real-life condition data, the employed feature extraction methods with their vector dimensions are as follows:

- MFCC_E: MFCC + Log-Energy (13)
- MFCC_EZ: MFCC_E with Cepstral Mean Normalization (CMN) (13)
- MFCC_ED: MFCC_E + delta (26)
- MFCC_EDZ: MFCC_E + delta with CMN (26)
- TEO_E: TEO-CB-Auto-Env + Log-Energy (17)

An analysis window of 32 msec duration is used with a 16 msec skip rate for 6-kHz speech data. For the MFCC, a standard algorithm suggested by the European Telecommunication Standards Institute (ETSI) was employed [8], where the 23-Mel-filterbank outputs are transformed to 12 cepstral coefficients adding log-energy (i.e., c1-c12, logE). For the TEO-based feature, 16 Gabor bandpass filters were used considering full bandwidth (i.e., 0-3kHz). Based on log-energy, speech frames were selected and then submitted to the angry speech classifier with silence segments dropped.

The results in Fig. 3 present the anger detection performance employing a "speaker independent" model for the "open-speaker" test as Equal Error Rate (EER, %). For the open-speaker test, when testing a speaker, data from all other remaining speakers were used for model training. All EERs in this paper represent an averaged value of EERs for 8, 16, 32, and 64 number of mixtures for the GMM classifier. In this open-speaker task, TEO_E outperformed MFCC_E and MFCC_EZ, and MFCC_EDZ showed the best performance as a single feature approach. Combination methods with TEO_E and MFCC_EDZ in both feature and decision levels gave improved performance compared to the single feature methods. From a series of performance evaluations, MFCC_EDZ showed consistently better performance compared to MFCC_ED in both single feature and combination approaches with a few outliers. The results show that CMN was effective at addressing channel distortion included in the speech samples recorded in telephone channel environments.

EERs in Fig. 4 present detection performance using a "speaker independent" model for the "closed-speaker" test. For the closed-speaker test, a set of a test speaker was split into two parts (e.g., part A and B). When testing with part A of a test speaker, data from all other remaining speakers and also part B of the test speaker were included for model training. Therefore, there is still no utterance (i.e., text) overlap between training and test data. Since the real-life data used in this study represents a limited amount, such a closed-speaker test would reflect the performance with more sufficient amount of training data, which could include a greater number of spectral patterns similar to the test speaker resulting in a more reliable acoustic model. Here, we also obtained a similar trend in EERs compared to the open-speaker test, showing that the combination methods with TEO_E and MFCC_EDZ are most effective at angry speech detection.

5.2. Performance Evaluation Employing Emotional Language Model Score

Here, we conducted performance evaluation employing the emotional language model score to leverage content struc-
Table 3. EERs of angry speech detection based on language model score (%).

<table>
<thead>
<tr>
<th></th>
<th>mix 2</th>
<th>mix 4</th>
<th>mix 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open-speaker test</td>
<td>45.33</td>
<td>46.18</td>
<td>44.08</td>
</tr>
<tr>
<td>Closed-speaker test</td>
<td>45.74</td>
<td>42.82</td>
<td>46.18</td>
</tr>
</tbody>
</table>

Fig. 5. EERs with employing EMLS for open-speaker test.

The results in Figs. 5 and 6 show angry speech detection performance when combining the classifiers based on acoustic features and the emotional language model score (ELMS). The likelihood scores of acoustic feature based and the ELMS based classifiers were combined with scale weights of 0.75 and 0.25 at the decision stage. It is quite interesting that there were consistently significant improvements for all feature evaluations for both open and closed speaker tests. By employing the emotional language model score, we obtained 33.05% and 29.87% in EERs for classifier combination of TEO_E and MFCC_EDZ (i.e., System II) for open and closed speaker test respectively.

6. CONCLUSIONS

This study has proposed an effective angry speech detection scheme for real conversational speech recorded over a telephone network. Analysis on the collected conversational speech shows a distinctive property in content structure between neutral and angry speech such as word distribution and frequently occurring words. In this study, a classifier based on emotional language model score was combined with acoustic feature based classifiers for angry speech detection. Experimental results demonstrated that the proposed method with the emotional language model score is more effective at increasing angry speech detection. We obtained up to a 6.23% decrease in EER by combining TEO-based feature, conventional MFCC, and emotional language model score, compared to single feature method for closed-speaker test (36.10% → 29.87%).

7. REFERENCES