Identification of four class emotion from Indonesian spoken language using acoustic and lexical features

Fatan Kasyidi and Dessi Puji Lestari
School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, Indonesia
23515025@std.stei.itb.ac.id, dessipuji@informatika.org

Abstract. One of the important aspects in human to human communication is to understand emotion of each party. Recently, interactions between human and computer continues to develop, especially affective interaction where emotion recognition is one of its important components. This paper presents our extended works on emotion recognition of Indonesian spoken language to identify four main class of emotions: Happy, Sad, Angry, and Contentment using combination of acoustic/prosodic features and lexical features. We construct emotion speech corpus from Indonesia television talk show where the situations are as close as possible to the natural situation. After constructing the emotion speech corpus, the acoustic/prosodic and lexical features are extracted to train the emotion model. We employ some machine learning algorithms such as Support Vector Machine (SVM), Naive Bayes, and Random Forest to get the best model. The experiment result of testing data shows that the best model has an F-measure score of 0.447 by using only the acoustic/prosodic feature and F-measure score of 0.488 by using both acoustic/prosodic and lexical features to recognize four class emotion using the SVM RBF Kernel.

1. Introduction
In recent time, research on emotion recognition for several languages has shown significant improvement. This helps to increase the quality of human-computer interaction to have more natural and affective communication. Speech emotion recognitions use many classification techniques such as Support Vector Machine, Naive Bayes, Random Forest, Gaussian Mixture Model (GMM) and Artificial Neural Network (ANN) [1–4]. In [2-3], GMM and ANN showed better performances comparing to the Random Forest and Naive Bayes. For Indonesian Language, research of emotion recognition from speech can be found in [1] where SVM-based model using the acoustic features has been developed. The model they developed shows accuracy of 68.31% for recognizing four class of emotions (angry, happiness, contentment, and sadness). They used only the acoustic/prosodic characteristics. Other research has also been established recently for Indonesian spoken language to recognize the positive and negative emotion using acoustic and lexical features [6].

As for the features, [2,3,6] used not only the acoustic/prosodic features, but also lexical features to improve performance of the system. Although, previous researches show that acoustic features are very dominant to be used for speech emotion recognition, adding lexical features also give promising results to increase performance of the system. Several type of acoustic features have been explored, including the prosodic features (such as pitch, energy, duration and higher order formants), spectral
features and cepstral features [1–3]. It includes the derivation of statistical function such as mean, standard deviation, maximum, minimum, range etc.

In this paper, we study four class emotion consist of angry emotion, happiness emotion, contentment, and sadness emotions as in [1]. However other than acoustic feature, we also employ the lexical features. In order to train the emotional model, the speech emotion corpus is collected from Indonesian TV Shows, Indonesia Lawyers Club (ILC), Mata Najwa, Tonight Show and Sarah Sechan. In order to extract the lexical features, we annotate the transcript manually by listening to the audio data. We use the Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) as has shown an effective performance of recognizing the emotion along with acoustic/prosodic feature [2].

The remainder of this paper is written as follows: the related works is explained in section 2. In section 3, the database collection is discussed. We explain the configuration of the experiment in section 4. In section 5, we describe the result of the experiment, and finally in Section 6 a conclusion and future works are presented.

2. Related Works
In recent years, some research for emotion recognition from spoken language have been established. A study for emotion recognition has been established in English [2]. They used the USC-IEMOCAP as the database that consists of 12 hours of audio-visual data, including speech and text transcription. Every utterance has an emotion label category such as angry, happy, etc and dimensional labels such as valence and activation. There are in total 5531 utterances and 419.4 minutes for the overall duration. The feature extraction process performed to get acoustic and lexical features from those utterances as feature representation. In this study, SVM linear kernel was utilized to model the emotion representation.

Other work on emotion recognition was performed in Chinese [3]. In this study, lexical feature along with acoustic/prosodic feature has been utilized to construct the emotion model. Acoustic/prosodic feature such as MFCC, jitter-related, shimmer related and emotional salient segment was included. For lexical feature, they use semantic labels as the lexical aspect using the HowNet. HowNet is an online common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of connoted in the lexicon containing Chinese and English equivalents. The accuracy of this model is 83.55%.

3. Data Collection
The availability of speech corpus containing emotions is unavoidable in order to develop emotion model developed using the machine learning technique. Thus, it is necessary to construct speech emotion corpus for Indonesian language since it is difficult to find publicly available corpus for this task. To construct the corpus, the corpus construction procedures from the previous work was followed [1]. In addition to that, we transcribe manually all utterances in the speech corpus to get their corresponding transcript. There are 3 main steps when collecting the speech corpus: data collection, utterance segmentation, and emotional and transcript labeling as can be seen in Figure 1.

Speech corpus is collected from four Indonesian TV talk show where there are many emotional dialog occur: Indonesian Lawyer club, Tonight Show, Mata Najwa, and Sarah Sechan. After conducting the segmentation process, we get 3652 utterances with total duration of recording is 5 hours 10 minutes and 40 seconds including the neutral emotion. However since the the neutral emotion utterance can bias the emotion recognition, we separate the utterances having neutral emotion from the dataset. Thus, after removing the neutral utterances, our data set contains only 1082 utterance with total duration of recording is 1 hour and 7 minutes. Each utterance contains one emotion to ensure the consistency and to avoid the change and transition of emotion for simplicity purpose. Before we annotate the emotion class, we transcribe each utterances manually to get their corresponding transcript. The transcribing process was conducted by 4 annotators that transcribe the speech utterances carefully by listening to each utterances in the database.
After that, every annotator was asked to give their judgment regarding the emotion label of each utterance. To evaluate and decide the emotion in each utterance, majority voting was employed to conclude the emotion class of each utterance. One of the observer is a psychologist that has more weight value than that of another observer. We use an expert decision if there is more than one decision with the same number of votes. In the end, we collect 362 utterances for angry, 262 utterances for happiness, 209 utterances for contentment and 149 utterances for sadness. The data for each class is not balanced since it is difficult to gather such data. Later on, to handle the imbalance data, we apply the resampling technique Synthetic Minority Oversampling Technique (SMOTE). In Figure 2, we can see clearly that the data is not balanced. We have angry as majority class and sadness as minority class.

![Figure 1. Corpus Construction Steps](image1)

![Figure 2. Data for Each Emotion Class](image2)
For conducting the experiments, the data that we have collected is divided into two sets: one for
development set and one for testing set. The development set contains 982 utterances while the testing
set contains 100 utterance with equal distribution between each class, i.e. 25 utterances per emotion
class.

4. Experimental Setup
For determining the best features set, we conduct three kind of model development: 1) train the model
by using only the acoustic features; 2) train the model by using lexical features; 3) train the model by
using both acoustic and lexical features. Those 3 features is employed to train the emotional model
using 3 classifiers: SVM, Naive Bayes, and Random Forest. Before we start the experiment, we
resampled the data to cope with the data imbalance so there are experiments with and without
sampling. We used 5-fold cross-validation and testing schema. The important thing that related to
the cross-validation is how to apply the SMOTE in that scheme. As we know that 5-fold cross validation
divides the data into two groups in every fold: training and testing set. The SMOTE method only uses
the training set in every fold to avoid the duplication and overfitting model in the result. Since the
angry class is the majority class, we did not over-sampled the data in this class, while for other three
emotion classes, we conduct the oversampling with sampling percentage as shown in Table 1.

Table 1. Percentages of SMOTE for emotion class

| Emotion    | Percentage (%) |
|------------|----------------|
| Angry      | 0              |
| Happiness  | 30             |
| Contentment| 50             |
| Sadness    | 120            |

In total, there are 40 experiments that combine features set, experimental schemes, and type of
classifier. We compared the performance of every model to find the highest accuracy model.

4.1. Acoustic/prosodic Feature
We extract acoustic/prosodic features for constructing the emotion model. Two acoustic/prosodic
features, the emo large feature set, the Low Level Descriptor of Interspeech 2010 set, and the
Interspeech 2010 Paralinguistic Challenge feature set are employed and compared as previous work of
detecting emotion from Indonesia spoken language used only the emo large feature and the Low Level
Descriptor of Interspeech 2010 [1]. The extracted feature for the emo large feature and the Low Level
Descriptor of Interspeech 2010 consists of 57 low-level descriptors (LLD) as shown in Table 2 and
Table 3.

Table 2. Low Level Descriptor of Emo Large [1]

| Cepstral features | Spectral features                      |
|-------------------|----------------------------------------|
| MFCC 0-12         | Mel-Spectrum bins 0-25                 |
|                   | Zero crossing rate                      |
|                   | 25% spectral roll-off points           |
|                   | 50% spectral roll-off points           |
|                   | 75% spectral roll-off points           |
|                   | 90% spectral roll-off points           |
|                   | Spectral flux                          |
|                   | Centroid                               |
|                   | Relative position of spectral maximum and minimum |
Table 3. Low Level Descriptor of Interspeech 2010 [1]

| Energy features | Voicing-related features |
|-----------------|--------------------------|
| Logarithmic energy | F0                       |
| Energy bands 0-250Hz | F0 envelope              |
| Energy bands 0-650Hz | Probability of voicing |
| Energy bands 250-650Hz |

The Low Level Descriptor of Interspeech 2010 Paralinguistic consists of 31 LLD and 1582 number of features [2]. This feature set has not been used in previous work [1]. However, recently this feature set has been used in Indonesia language emotion recognition to recognize 7 classes of emotions as can be found in [7]. The 31 LLD features set can be seen in Table 4.

Table 4. Low Level Descriptor of Interspeech 2010 Paralinguistic [2]

| LLD                  |
|----------------------|
| Continuous Qualitative Cepstral |
| Energy: loudness     | Local jitter          | MFCC 0-14        |
| Pitch: F0 final      | JitterDDP             | LogMelFreqBand 0-7|
| F0 final envelope    | Local shimmer         |                  |
| Line Spectral Pair Frequencies | Voicing final unclipped |

4.2. Lexical Feature

We compute Bag-of-Word (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) to extract the lexical features from each transcript of the speech utterance. The BoW method count the frequency of a word in the document without considering the word order, semantic structure and the grammar [2]. The vector representation of BoW is a normally used feature representation for a text document. The TF-IDF is used to evaluate how important a word is to the certain document in the corpus. The TF-IDF weight of term i in document j as calculated using formula (1).

\[ w_{i,j} = t_{f_{i,j}} \times \log \frac{N}{d_{f_{i}}} \]  

In this study, we perform several preprocessing step to the transcription before BoW and TF-IDF are being constructed. We used two preprocessing: formalization and stop-word removal.

5. Result

In this section, the result of all experiments are described. To construct the models, 982 speech utterances of development dataset are being used. We calculate the average of F-measure in every model. As we mention before, there are three models having different feature set: the acoustic features set, lexical features set, and acoustic lexical (AL) feature set. Each features set will be used to train emotional model using the Naive Bayes, Random Forest, and SVM with Radial Basis Function (RBF) kernel and linear kernel. The result of the acoustic model can be seen Table 5.
Table 5. Result of Experiment for Acoustic Model (Two First Column using the Low Level Descriptor of Interspeech 2010 Paralinguistic and the last column using the Emo Large Feature Set)

| Model         | Without Resampling F-Measure | Resampling with SMOTE F-Measure | Emo Large Feature Set F-Measure |
|---------------|-----------------------------|---------------------------------|--------------------------------|
| Naive Bayes   | 0.377                       | 0.393                           | -                              |
| Random Forest | **0.389**                   | **0.409**                       | -                              |
| SVM RBF       | 0.378                       | 0.400                           | 0.411                          |
| SVM Linear    | 0.388                       | 0.340                           | -                              |

It is shown in Table 5 that when using the Low Level Descriptor of Interspeech 2010 Paralinguistic acoustic feature set, the Random Forest (RF) is the most optimal model for acoustic model with SMOTE and the detail of F-measure for each emotion can be seen in Table 6. It can be seen from Table 6 that it is easier to identify angry and happiness emotions (emotion with more active energy), rather than recognizing sad and contentment emotions.

Table 6. F-measure of Acoustic Model using RF

| Emotion      | F-Measure |
|--------------|-----------|
| Angry        | 0.535     |
| Happiness    | 0.449     |
| Contentment  | 0.264     |
| Sadness      | 0.237     |

To see the effectiveness of using the lexical features, we used the lexical features only and the result of using BoW and TF-IDF can be seen in Table 7.

Table 7. Result of Experiment for Lexical Model

| Model         | Without Resampling BoW | Without Resampling TF-IDF | Resampling with SMOTE BoW | Resampling with SMOTE TF-IDF |
|---------------|------------------------|---------------------------|---------------------------|-----------------------------|
| Naive Bayes   | 0.359                  | 0.339                     | 0.366                     | 0.382                       |
| Random Forest | 0.399                  | 0.375                     | 0.383                     | **0.404**                   |
| SVM RBF       | 0.199                  | 0.198                     | 0.199                     | 0.199                       |
| SVM Linear    | 0.394                  | 0.364                     | 0.379                     | 0.341                       |

It is shown in Table 7 that Random Forest is the most optimal model for lexical model TF-IDF with SMOTE and the detail of F-measure for each emotion can be seen in Table 8.

Table 8. F-measure of Lexical Model TF-IDF using RF

| Emotion      | F-Measure |
|--------------|-----------|
| Angry        | 0.494     |
| Happiness    | 0.399     |
| Sadness      | 0.224     |
| Contentment  | 0.381     |
Finally, we combine the acoustic and lexical features and the results can be seen in Table 9.

| Model         | Without Resampling | Resampling with SMOTE | Emo Large Feature Set |
|---------------|--------------------|-----------------------|-----------------------|
|               | F-measure          | F-measure             | F-measure             |
| Naive Bayes   | BoW: 0.380         | TF-IDF: 0.375         | BoW: 0.397            | TF-IDF: 0.400             | -                        | -                        |
| Random Forest | 0.380              | 0.380                 | 0.400                 | 0.394                     | -                        | -                        |
| SVM RBF       | 0.388              | 0.389                 | 0.402                 | 0.374                     | 0.380                    |
| SVM Linear    | 0.336              | 0.335                 | 0.325                 | 0.329                     | -                        | -                        |

It is shown in Table 9 that SVM RBF kernel is the most optimal model for acoustic lexical (AL) model BoW with SMOTE and the detail of F-measure for each emotion can be seen in Table 10.

| Emotion | F-Measure |
|---------|-----------|
| Angry   | 0.559     |
| Happiness | 0.377 |
| Contentment | 0.257 |
| Sadness | 0.273     |

The result from the experiment of development data modeling shows that F-measure for each optimal model consecutively are 0.447, 0.404 and 0.408. After that, we construct the testing scenario using testing dataset consist of 100 utterances with 25 utterances for each emotion. Testing scenario constructed only for the most optimal model from 3 global model (acoustic, lexical and AL). The result can be seen in Table 11.

| Emotion | F-measure |
|---------|-----------|
| Angry   | 0.639     |
| Happiness | 0.471 |
| Contentment | 0.293 |
| Sadness | 0.392     |
| Average | 0.447     |

According to Table 11, the emo large as the configuration used in the previous work [1] have a lower accuracy than the acoustic lexical features (AL). This result shows that lexical features have an effect to enhance the accuracy of emotion recognition. Although, the lexical features can improve the recognition result, but it is not always the case. The data with id file P001111 is the example of inconsistency in recognizing the emotion. The AL model could not improved the ability to recognizing the emotion. But there are some examples where the AL model can enhance and fix the recognition such as can be seen in the data with id file P018015. Both of these examples can be seen in Table 12.
Table 12. Result of Prediction

| File | Transcript | Label  | Acoustic Model | Lexical Model | AL Model |
|------|------------|--------|----------------|---------------|----------|
| P001111 | selamat berpromo juga semoga di album kedua ini penjualan albumnya meningkat kembali | happiness angry | happiness | angry |
| P018015 | sebegitu dekatnya yang bikin aku makin jatuh cinta | happiness sadness | happiness | happiness |

Human character and psychology is also affects the recognition. Ethnic in a certain area usually affect the character, especially how the way someone talking. For example, there are some people that talk in a relatively high pitch, even though they are not angry. Thus, it make the recognition is more difficult.

6. Conclusion
In this paper, we have extended the work on emotion recognition in Indonesian spoken language from our previous works. For this study, we constructed emotional corpus in Indonesian spoken language from Indonesian television talk show which provides the natural emotional data for the corpus. By using this corpus, we construct three emotion recognition model using acoustic: acoustic by using both Emo Large Feature Set that has been used in the previous research and the Interspeech 2010 Feature Set, lexical features using the Bag of Word and TF-IDF, and the lexical and Acoustic Lexical (AL) model. Experimental results show that the best model is achieved when using both acoustic (the interspeech 2010 feature set) and lexical features (BoW) with 0.488 of F-measure score using the SVM RBF Kernel.

References
[1] Lubis N, Lestari D, Purwarianti A, Sakti S, Nakamura S. Emotion recognition on Indonesian television talk shows. In: 2014 IEEE Spoken Language Technology Workshop (SLT). 2014. p. 466–71.
[2] Jin Q, Li C, Chen S, Wu H. Speech emotion recognition with acoustic and lexical features. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. p. 4749–53.
[3] Wu CH, Liang WB. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels (Extended abstract). In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). 2015. p. 477–83.
[4] Noroozi F, Akrami N, Anbarjafari G. Speech-based emotion recognition and next reaction prediction. In: 2017 25th Signal Processing and Communications Applications Conference (SIU). 2017. p. 1–4.
[5] Bhakre SK, Bang A. Emotion recognition on the basis of audio signal using Naive Bayes classifier. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2016. p. 2363–7.
[6] Kurniawati P, Lestari D. Positive and Negative Emotions Identification from Speech Signal using Acoustic and Lexical Features. In: 2017 International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA). 2017.
[7] Kurniawati P, Lestari D, Khodra M. Speech Emotion Recognition From Indonesian Spoken Language Using Acoustic and Lexical Features. In: The 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment. Seoul, Korea; 2017.
[8] Elhadad MK, Badran K, Salama GI. A novel approach for ontology-based dimensionality reduction for web text document classification. In: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). 2017. p. 373–8.