Recognition and Analysis of Emotion in Indonesian Conversational Speech

Nurul Lubis[1,2], Sakriani Sakti[1], Graham Neubig[1], Tomoki Toda[1],
Dessi Lestari[2], Ayu Purwarianti[2], and Satoshi Nakamura[1]

[1] Nara Institute of Science and Technology
[2] Institut Teknologi Bandung

Abstract: The importance of incorporating emotional aspect in human computer interaction continues to arise. Unfortunately, exploration of the subject in Indonesian is still very lacking. This paper presents the first study of emotion recognition in Indonesian on conversational speech. We construct our corpus, IDESC, by making use of television talk show recordings in various topics of discussion, yielding colorful emotional utterances. Using the corpus, we then build a support vector machine (SVM) that classifies Indonesian speech in terms of emotion based on its acoustic features. We perform feature selection and parameter optimization while building the classifier to optimize the recognition performance, resulting in absolute 11.9% increase of accuracy. Lastly, we perform analyses on our corpus and evaluation result to gain better insight of emotion occurence in Indonesian speech.

1 Introduction

Human computer interaction technologies aim at the most natural form of interaction possible by matching that of human and human. In that sense, the interaction should not only focus on completion of certain tasks, but also engagement with user on an emotional level. This requires a set of capabilities in a machine; recognizing, interpreting, processing, and simulating human affects.

To examine different sides and angles of this problem a number of emotional challenges have been held from year to year, e.g INTERSPEECH [1] [2] and AVEC [3] [4]. Along with research and studies in this field, interaction between human and computer have advanced to better facilitate the emotional aspect of the interaction.

For Asian languages, there exist a number of studies and findings related to emotion in computing. For Chinese, researchers have studied the effect of switching the stimulus in user, involving affective systems [5]. In Tagalog, an automated narrative storyteller was constructed with an average precision of 86.75% in expressing a particular emotion [6]. Unfortunately, in Indonesian, research on emotion recognition is non-existent—even the resource to conduct studies and research on is still very lacking.

This paper presents the first study of emotion recognition in Indonesian. We construct a speech corpus from television talk show recordings in various topics of discussion, yielding colorful emotional utterances. Utilizing the corpus, we then train and evaluate our emotion classifier. We observe and analyze the training and evaluation process to have better insight of emotion in Indonesian speech.

2 Previous Works

One of the early studies on speech based emotion recognition is performed on acted utterances in the English language [7]. The study reports a novel approach for classifying speech based on its emotion content and the promising acoustic features for the task. More recently, real-time recognition have been constructed using acted emotion corpus intended for teaching autistic children about simple and complex emotions [8].

Emotion recognition has been applied in spoken dialogue systems to deliver a more natural experience to user. This includes studies on emotion triggers on human spoken dialogue [9] and generation of emotionally coloured conversation [10]. In this context, spontaneous, naturalistic, or induced emotional speech data is preferable as they better mimic interaction between human and computer.

Despite the number of research in emotion recognition, only a few utilize human-human conversation that really mimics natural interaction—most make use of speech data recorded in specific situation, guidelines, or scenario, thus compromising the nature and emotion range of the resulting data. In this study, we utilize real human conversations to ensure applicability to actual human-computer interaction.

3 Corpus Construction

Construction of our corpus, the Indonesian Emotional Speech Corpus (IDESC), comprises three steps. The first is data collection, during which contents for the corpus are gathered. After collection, the content is then segmented into speech utterances. Each segment is then annotated or labeled based on its emotion content. The overall construction process of the corpus is demonstrated in Fig. 1. Each of these steps will be explained in detail in this section.

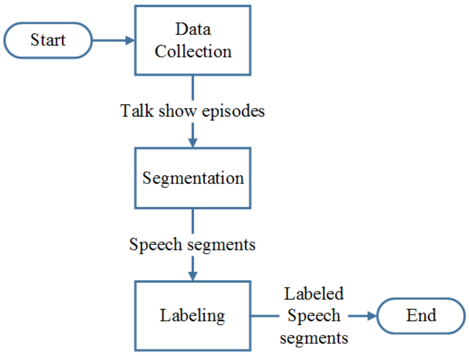


図 1: The steps of corpus construction

3.1 Data Collection

We collect the speech data in Indonesian from various television talk shows. Television talk shows provide clean speech recordings with distinguishable dialogue turns, resulting in good quality speech data. The dialogue format gives us natural speech utterances. Furthermore, with different guests in each episode, speech data from a number of speakers can be obtained.

We select three episodes from different talk shows to cover broader range of emotions. The first show is ”Mata Najwa”, with discussions focusing on politic related subjects. The second show is ”Kick Andy”, with topics in the area of humanity. The third show is ”Just Alvin”, with focus in celebrities, entertainment, and lifestyle. The different topics are expected to provide more varied emotion content in the collected data.

In total, the three talk show episodes are 2 hours, 25 minutes, and 39 seconds in length. Audio is available at 16 kHz and 16 bits per sample. There are 18 speakers in total; 12 male speakers and 6 female.

3.2 Segmentation

The collected data is segmented into speech utterances. We segment the speech manually to ensure quality, as segmentation using an existing automatic speech recognizer may introduce errors to the result. During the process, we make sure the emotion content is consistent for each segment. This is done so that the resulting segments are relevant in emotion recognition. However, this doesn’t limit other approach of segmentation in the corpus to suit other task in advanced human-computer interaction.

Segmentation is done using speech processing tool Audacity.[1]. In total we obtained 2179 speech segments worth 1 hour, 34 minutes, and 49.7 seconds in length.

3.3 Labeling

We label the segments manually based on human recognition. 5 human labelers are employed, 3 females and 2 males. Before labeling, the labelers are briefed regarding the corpus and the objective of the task. We define 5 emotion labels derived from the widely-known valence-arousal scale: neutral, happiness, anger, sadness, and contentment. These classes are general, yet it covers all emotions in daily human interactions. This set of emotion labels gives a good foundation in further development of IDESC, where more specific emotion terms can be defined.

A segment is labeled neutral if not enough affect is detected in the speech. If a speech segment shows active expression of a positive emotion, it is labeled happiness; if it’s of negative emotion, it’s labeled sadness. Passive expression of positive emotion yields the contentment label, and on negative emotion, sadness. This labeling rule is simple and straightforward,

[1] http://audacity.sourceforge.net/

as we want to start with a less complicated task and progress as we develop IDESC. After all segments are labeled, we obtain the finished emotional corpus in Indonesian.

4 Experiment Set Up

In training our SVM, we employ the widely used library for support vector machines, libSVM [11] using the Radial Basis Function (RBF) kernel. Fig. 2 gives an overview of the emotion recognition model construction, followed with evaluation. The process comprises scaling, feature selection, and parameter optimization to obtain a model that is as good as possible. These steps are recommended in [12] and have been applied in similar tasks by [8].

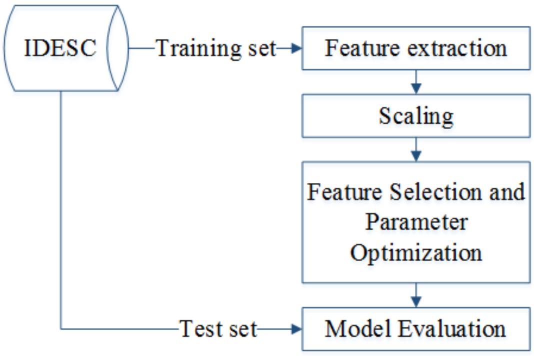


図 2: Experiment and evaluation flow

In the experiment, we exclude the segments with a neutral label to keep our focus in recognizing present emotion. We part our IDESC into training set and test set with a 85:15 ratio. We use the training set to construct our emotion recognizer and the test set to subsequently evaluate it. We obtain 1155 segments in the training set with a distribution as follows: 204 segments labeled as happiness, 467 as anger, 228 as sadness, and 459 as contentment.

On the training set, we extract acoustic features as the basis of emotion recognition. For reproducibility, we employ an open-source feature extraction toolkit openSMILE [13]. The employed feature set is the official openSMILE `emo_large.conf` feature set. We choose a large feature set to extract as much relevant information as possible from speech to be filtered by feature selection in the next step.

The feature set includes detailed statistical description of the basic speech features with many spectral and further descriptors. The low level descriptors (LLD) are listed in Table 1, categorized into cepstral, spectral, energy, and voicing-related features. For each LLD as well as its delta and acceleration coefficients, 39 statistical functionals are computed. These functionals include values such as mean, standard deviation, percentiles and quartiles, linear regression functionals, or local minima/maxima related functionals.

Cepstral features (13)
MFCC 0-12
Spectral features (35)
Mel-Spectrum bins 0–25, zero crossing rate, 25%, 50%, 75%, and 90% spectral roll-off points, spectral flux, centroid, relative position of spectral maximum and minimum
Energy features (6)
logarithmic energy, energy in bands from 0–250 Hz, 0–650 Hz, 250–650 Hz, 1–4 kHz, 3010–9123 Hz
Voicing-related features(3)
F0 (subharmonic summation (SHS) followed by Viterbi smoothing), F0 envelope, probability of voicing

表 1: Low level descriptors of the feature set

For each segment, we scale the value of the extracted features to a [-1, +1] range to avoid overpowering of features of bigger value and numerical difficulties in further SVM computation. After scaling the features, we perform feature selection based on F-score to eliminate features that are possibly irrelevant or causing noise in the training data.

First, we calculate the F-score of each extracted feature and sort them in descending order and exclude features with F-score of 0. By continually dividing the number of features by 2, we loop to experiment with different numbers of top scoring features. On each loop, we obtain the average recognition rate by performing 5-fold cross validation. The subset with the best validation rate is chosen as the optimal feature set.

During feature selection, we also perform parameter optimization using the grid search algorithm while evaluating each feature subset. This is done in a brute-force manner by testing pairs of learning parameters cost and gamma $(C, \gamma)|C \in \{2^{-5}, 2^{-2}, ..., 2^{13}, 2^{15}\}, \gamma \in \{2^{-15}, 2^{-13}, ..., 2, 2^{3}\}$ and choose the pair with the best 5-fold-cross-validation rate. Based on the experiment, we build our emotion classifier accordingly.

5 Experiment Result and Analysis

In this section we lay out detailed analysis on our corpus construction and experiment described in previous sections. First, we analyze the emotion content on IDESC to get better insight of the corpus. Second, we evaluate our model by performing emotion recognition on our test set and subsequently analyze the result. Third, we try to further investigate the effect of learning optimization in our experiments.

5.1 Data Analysis

We analyze the emotion content of our constructed speech corpus, IDESC. Firstly, we analyze the agreement level between human annotators using Fleiss' Kappa. These numbers are presented in Fig. 3. The statistic shows that the annotators have the highest agreement on labeling segments with anger emotion, with moderate agreement at 0.55, while lowest on neutral emotion, with slight agreement at 0.28. Anger ranked first on annotators' agreement level, followed by happiness, contentment, sadness, and neutral. This tells us that active emotions are more uniformly recognizable than passive ones, or the absence of emotion. The overall agreement level of the annotators on the entire corpus reaches 0.23, which can be interpreted as slight agreement.

Secondly, from each talk show, as well as IDESC in overall, we visualize its distribution of emotion labels. We detect different tendency of emotion occurrences in each talk show. We then try to draw correlation of this tendency to the topic of the discussion. The distribution of the emotion class in IDESC is visualized in Fig. 4.

In general, neutral and passive-positive emotion seem to be the most common occurrence in the collected dialogues. This is expected, as talk shows are rather formal in format and broadcasted to a large amount of audience. However, at certain parts, different emotions do naturally occur as the result of the topic discussed. This correlation between topic of discussion and emotion occurrence will be beneficial in further data collection of certain emotion content.

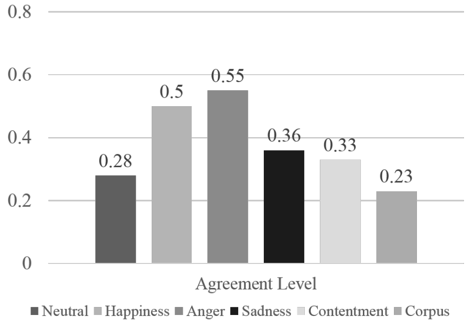


図 3: Agreement level between annotators measured with Fleiss' Kappa

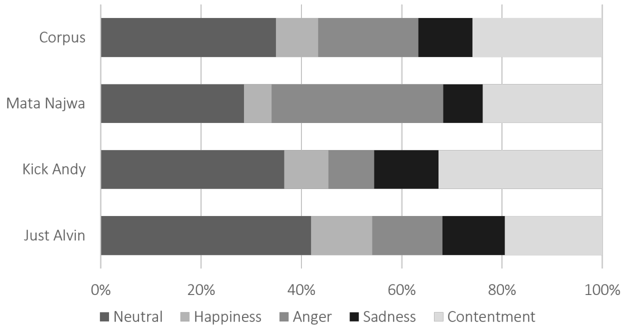


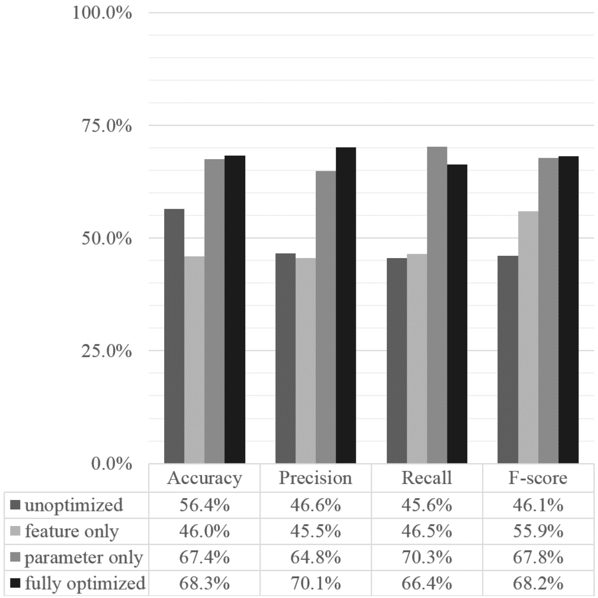
図 4: Distribution of emotion labels in the corpus

We analyze the emotion content (aside from neutral) of each talk show and correlate it to its main topic. "Mata Najwa", which provides a discussion centering in politics, is the talk show with most positive-negative emotion occurrences. On the other hand, the story telling of enriching life experiences in "Kick Andy" gives us the many passive-positive occurrences. Meanwhile, "Just Alvin" with its focus in entertainment and celebrities, seem to have balanced occurrences on the four labels of emotion. Overall, we obtain emotion label distribution as shown in the first bar of Fig. 4. The composition is slightly imbalanced, with happiness and sadness as the least occurring emotions.

5.2 Feature Selection and Parameter Optimization

We evaluate our learning optimization algorithm explained in Sec. 4. We build 4 models with different learning optimizations and perform recognition of our test data on them. The test set consists of 202 speech segments in Indonesian with emotion distribution as follows: 32 segments labeled as happiness, 71 as anger, 33 as sadness, and 66 as contentment.

The first model is build with no learning optimization. On the second model, we apply only feature selection and on the third, parameter optimization. The fourth model is built with both feature selection and parameter optimization. Based on the recognition, we calculate 4 performance measures as comparison of the models: accuracy, precision, recall, and F-score. The unoptimized model serves as the baseline. These numbers are presented in Fig.5.



	Accuracy	Precision	Recall	F-score
■ unoptimized	56.4%	46.6%	45.6%	46.1%
■ feature only	46.0%	45.5%	46.5%	55.9%
■ parameter only	67.4%	64.8%	70.3%	67.8%
■ fully optimized	68.3%	70.1%	66.4%	68.2%

図 5: Comparison of model performance with different learning optimization

Our figure shows that with feature selection only, we're able to obtain a model with higher F-score than unoptimized model, but with lower accuracy. On the other hand, parameter optimization alone is able to improve recognition performance on all measurements. This tells us that while selecting the proper features is important in building a classifier, to fit the model to the problem is just as essential if not more. Combining the two techniques gives us the fully optimized model with the best scores on almost all performance measures compared to the rest of the models.

		Prediction			
		1	2	3	4
Label	1 (*happiness*)	**18**	2	1	11
	2 (*anger*)	1	**55**	6	9
	3 (*sadness*)	0	1	**22**	10
	4 (*contentment*)	2	7	14	**43**

表 2: Confusion matrix for prediction of test set

		Hap	*Ang*	*Sad*	*Con*
One-against-one	*Happiness*				
	Anger	**94.17%**			
	Sadness	92.30%	86.53%		
	Contentment	83.67%	83.94%	72.72%	
One-against-all		88.18%	**88.61%**	86.63%	71.78%

表 3: One-against-one and one-against-all emotion recognition accuracy

5.3 Emotion Recognition

To analyse emotion in Indonesian speech further, we take a closer look on recognition by our fully optimized model, the proposed model in our study. Confusion matrix of the recognition by the model is presented in Table 2, with correct prediction marked with boldface.

Most incorrect predictions happen when a segment is or should be predicted as contentment. On the other hand, anger is the best recognized emotion. This demonstrates prominence of acoustic characteristics of the emotions; with contentment being the least prominent and anger being the most. It's worth noting that both emotions are of different valence and arousal range.

Furthermore, Table 3 presents recognition accuracy for the one-against-one and one-against-all scenarios, with the highest performance boldfaced and the lowest underlined. Classification between happiness and anger emotion yields the best accuracy for one-against-one classification, while sadness and contentment yields the lowest. On one-against-all, the best accuracy is obtained while classifying anger emotion from the rest, and the least, contentment.

6 Conclusion

We present a study on emotion recognition in Indonesian. We build an emotion recognizer from an emotionally colorful speech corpus in Indonesian, IDESC. In constructing our emotion recognizer, we attempt to obtain the best resulting model possible by optimizing the learning process with feature selection and parameter optimization. As a result, we achieve multiclass classification accuracy of 68.31% for 4 emotion classes. We perform analysis on our data as well as the experiment process and result to get more insight of emotion in Indonesian specch. This information is highly beneficial for further development in the language.

The overall result of this study is widely open for improvements. More data in Indonesian can be obtained in answering the scarcity of resource. Furthermore, a more complex emotion structure can be defined to obtain a model capable of recognizing more specific emotions. We look forward to get meaningful comparison to this work by testing different techniques and approaches to our corpus.

参考文献

[1] Bjoern Schuller, Stefan Steidl, and Anton Batliner, “The INTERSPEECH 2009 emotion challenge.,” in *INTERSPEECH*. Citeseer, 2009, vol. 2009, pp. 312–315.

[2] Bjoern Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Mueller, and Shrikanth S Narayanan, “The INTERSPEECH 2010 paralinguistic challenge.,” in *INTERSPEECH*, 2010, pp. 2794–2797.

[3] Bjoern Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic, “Avec 2011–the first international audio/visual emotion challenge,” in *Affective Computing and Intelligent Interaction*, pp. 415–424. Springer, 2011.

[4] Bjoern Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic, “Avec 2012: the continuous audio/visual emotion challenge,” in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 449–456.

[5] Liusheng Wang and Jing Li, “The switching effect involving the affective system in Chinese affective concept processing,” *Universal Journal of Psychology*, vol. 2, no. 5, pp. 151–157, 2014.

[6] John Christopher P. Gonzaga, Jemimah A. Seguerra, Jhonnel A Turingan, Mel Patrick A. Ulit, and Ria A. Sagum, “Emotional techy basyang: An automated filipino narrative storyteller,” *International Journal of Future Computer and Communication*, vol. 3, pp. 271–274, August 2014.

[7] Frank Dellaert, Thomas Polzin, and Alex Waibel, “Recognizing emotion in speech,” in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. IEEE, 1996, vol. 3, pp. 1970–1973.

[8] Tomas Pfister and Peter Robinson, “Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis,” *Affective Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 66–78, 2011.

[9] Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, Ayu Purwarianti, and Satoshi Nakamura, “Emotion and its triggers in human spoken dialogue: Recognition and analysis,” in *Proceedings of International Workshop on Spoken Dialogue Systems*, 2014, pp. 224–229.

[10] Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amier, and DKJ Heylen, “The sensitive artificial listner: an induction technique for generating emotionally coloured conversation,” 2008.

[11] Chih-Chung Chang and Chih-Jen Lin, “LibSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[12] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al., “A practical guide to support vector classification,” 2003.

[13] Florian Eyben, Martin Woellmer, and Bjoern Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.