EMOTION RECOGNITION ON INDONESIAN TELEVISION TALK SHOWS

Nurul Lubis[1,2], *Dessi Lestari*[1], *Ayu Purwarianti*[1], *Sakriani Sakti*[2], *Satoshi Nakamura*[2]

[1]School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia
[2]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
13510012@std.stei.itb.ac.id, {dessipuji,ayu}@stei.itb.ac.id, {ssakti,s-nakamura}@is.naist.jp

ABSTRACT

As interaction between human and computer continues to develop to the most natural form possible, it becomes more and more urgent to incorporate emotion in the equation. The field continues to develop, yet exploration of the subject in Indonesian is still very lacking. This paper presents the first study of emotion recognition in Indonesian, including the construction of the first emotionally colored speech corpus in the language, and the building of an emotion classifier through an optimized machine learning process. We construct our corpus using television talk show recordings in various topics of discussion, yielding colorful emotional utterances. In our machine learning experiment, we employ the support vector machine (SVM) algorithm with feature selection and parameter optimization to ensure the best resulting model possible. Evaluation of the experiment result shows recognition accuracy of 68.31% at best.

Index Terms— Indonesian, emotion recognition, speech, acoustic, SVM

1. INTRODUCTION

Over the years, interaction between human and computer continues to change to the better replicate interaction between humans. As emotion plays an important role in making a natural interaction, researches then put more effort in incorporating human emotion in the human-computer interaction. This necessity builds up the affective computing field, a field that studies and develops systems capable of recognizing, interpreting, processing, and simulating human affects.

A number of emotional challenges have been held from year to year to address various issues in the field. In 2009, INTERSPEECH tried to bridge the gaps between excellent research on human emotion recognition from speech and low compatibility of results [1]. They continued to address affective issues in 2010 through one of their sub-challenges [2]. In 2011, Audio Visual Emotion Challenge (AVEC) was held for the first time, aiming at multimedia processing and machine learning methods for automatic emotion analysis [3]. After that, AVEC 2012 tried to analyze emotion from its dimensions rather than identifying it as discrete states [4].

Spoken language technology in Indonesian starts to actively developed in the past few years. Currently, there exist a number studies on automatic speech recognition [5] [6], machine translation [7] [8], and speech synthesis [9]. Unfortunately, research on emotion recognition is non-existent—even the resource to conduct studies and research on is still very lacking. This is the reason we initiate the study of speech-based emotion recognition in Indonesian, starting with corpus construction, followed by the emotion recognizer.

This paper presents the first study of emotion recognition in Indonesian, including the construction of the first emotionally colored speech corpus in the language, and the building of an emotion classifier through an optimized machine learning process. We construct our corpus using television talk show recordings in various topics of discussion, yielding colorful emotional utterances. In our machine learning experiment, we employ the SVM algorithm with feature selection and parameter optimization to ensure the best resulting model possible.

The remainder of this paper is as follows. In section 2, we discuss previous research and studies relevant to the task at hand. We describe the construction of the corpus used in this study in section 3. Section 4 explains the configurations of our experiments. In section 5, we analyze out data, describe the result of our experiment, evaluate our emotion recognizer, and perform analysis on the result. Section 6 concludes the paper.

2. PREVIOUS WORKS

One of the early studies on speech based emotion recognition is performed on acted utterances in the English language [10]. The study reports a novel approach for classifying speech based on its emotion content and the promising acoustic features for the task. Over time, awareness of the importance of affect in human-computer interactions increases, setting off studies on affective computing for various languages in different contexts.

More specifically on speech based emotion recognition, in English, real-time recognition have been constructed using acted emotion corpus intended for teaching autistic children about simple and complex emotions. In German, a study

978-1-4799-7129-9/14/$31.00 ©2014 IEEE

exists using the combination of acted and spontaneous emotional speech [11], and in French using collection of dialogues in a customer service station [12]. Language wise, a study of English-German multi-lingual emotion recognition using interactive voice response (IVR) recordings argues that for different languages, different sets of features are needed to obtain optimal recognition performance [13].

Besides different languages, studies on emotion recognition have also been done for various purposes. Among others, researchers have tried to apply emotion recognition in spoken dialogue systems to deliver a more natural experience to user. This includes, but not limited to, studies on emotion triggers on human spoken dialogue [14] and generation of emotionally coloured conversation [15]. In this context, spontaneous, naturalistic, or induced emotional speech data is preferable as they better mimic interaction between human and computer.

Despite the research progress in emotion recognition, only a few number of researches utilize human-human conversation that really mimics daily interaction between humans—most make use of speech data recorded in specific situation, guidelines, or scenario, thus compromising the nature and emotion range of the resulting data. Moreover, currently, there is yet a research on emotion recognition for Indonesian. This is our motivation in building the first Indonesian emotion recognizer using real human conversations from television talk shows.

3. IDESC: INDONESIAN EMOTIONAL SPEECH CORPUS

In this section we explain in detail the construction of emotional speech corpus in Indonesian, containing only naturalistic emotion occurrences, as we aim at applicability to human-computer interaction technologies. Construction of 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.

3.1. Data Collection

We collect the speech data in Indonesian from various television talk shows. This approach is done before on the ”Vera am Mittag” corpus [16]. Television talk shows provide clean speech recordings with distinguishable dialogue turns, resulting in good quality speech data. The format gives us natural speech utterances. This will be beneficial in application for human-computer interactions, as interactions in that context happen naturally instead of acted. Television talk shows also provide speaker variance. With different guests in each episode, we're able to gather speech data from a number of speakers.

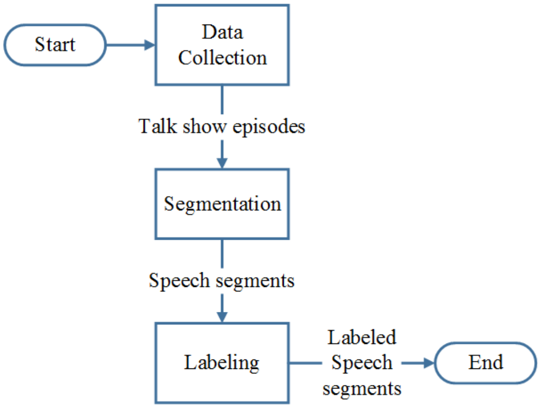

Fig. 1. The steps of corpus construction

We select three episodes from different kinds of talk shows to cover a broader range of emotions. The selected talk shows are very popular in the country, with discussions on engaging and interesting topics that trigger various emotions from the speakers. 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 a lighter focus in celebrities, their career, and life. 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. Video recordings of the show is obtained, but stripped down to audio only as we're currently focusing on speech data. Audio is available at 16 kHz and 16 bits per sample. There are 18 speakers in total; 12 male speakers and 6 female. This offers speaker variance in the corpus, even though the number of speech per speaker is not evenly distributed due to the role of each speaker in the talk show.

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. In other words, we avoid changes or transitions of emotion in a 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] As well as the segments, the segmentation is also provided in the form of time marking annotation of the start and

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

the end of the segments. 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 references are briefed regarding the objective of the labeling task and the corpus. We defined 5 emotion labels: 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, 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 [17] using the Radial Basis Function (RBF) kernel. Fig. 2 gives an overview of the emotion recognition model construction followed by 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 [18] and have been applied in similar tasks by [19].

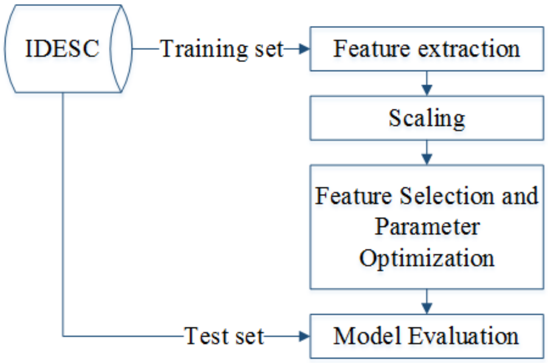

Fig. 2. Experiment and evaluation flow

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.

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

Table 1. Low level descriptors of the feature set

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 [20]. 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.

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 take a deeper look on the optimization result in machine learning as it yields information of emotional speech characteristics in Indonesian. Third, we evaluate our model by performing emotion recognition on our test set and subsequently analyze the result.

5.1. Data Analysis

We analyze the emotion content of our constructed speech corpus, IDESC. 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. 3.

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.

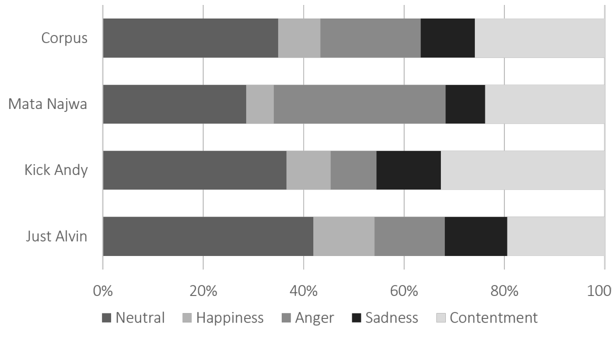

Fig. 3. Distribution of emotion labels in the corpus

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. 3. The composition is slightly imbalanced, with happiness and sadness as the least occurring emotions.

5.2. Model Optimization

During the experiment, we perform feature selection and parameter optimization to obtain the best model possible as described in detail previously in Sec.4. Fig.4 shows the result of this experiment, with y-axis showing the average recognition rate of different number of features in x-axis. We obtain best cross validation performance at 59.85% using 3312 best scoring features, with (8,0.001953) as learning parameters.

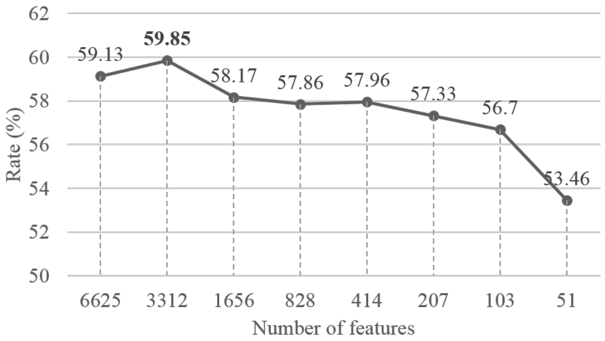

Fig. 4. Feature selection result

We take a closer look at the composition of optimal feature set obtained from the feature selection. The optimal set gives us a better idea of the feature types that best catches the characteristic of emotion in an Indonesian speech. The composition in the optimal set indicates that spectral features play a highly important role in Indonesian, composing 64.2% of the feature set, followed by cepstral features with 19%, energy with 10.8%, and voicing-related with 5%.

5.3. Emotion Recognition

We evaluate our model by performing emotion recognition on our test set in 3 scenarios: one-against-one, one-against-all, and multiclass classification. On one-against-one classification, we perform recognition per two emotion labels, covering all combinations possible. One-against-all scenario yields binary classification, where we test a certain label against the

		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%

Table 2. One-against-one and one-against-all emotion recognition accuracy

		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**

Table 3. Confusion matrix for prediction of test set

rest e.g. for anger one-against-all classification, happiness, sadness, and contentment are considered as not-anger. 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 results will be explained in this section.

Table 2 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.

A pattern can be observed from both scenarios. Recognition task that involves anger emotion is performed better, while one that involves contentment has relatively lower performance. It is also observed that recognition between passive emotions is more difficult to perform compared to that of active emotions.

Multiclass prediction from the model is presented in Table 3, with correct prediction marked with boldface. In multiclass classification, 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 finding is similar to that in one-against-one and one-against-all scenario. 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.

We calculate 4 performance measures according to the confusion matrix on Table 3: accuracy, recall, precision, and F-score. Our model achieves 68.31% accuracy, 66.38% recall rate, precision of 70.10%, and an F-score of 68.19%. Given the study is the first in the language and thus preliminary, we believe that these numbers are promising and up for improvement in further development of the recognizer.

6. CONCLUSION

We present the first study on emotion recognition in Indonesian. We construct an SVM based emotion recognizers using the first emotionally colorful speech corpus in Indonesian, IDESC. IDESC utilizes television talk show recordings in providing natural emotional speech covering a broad range of emotions. In constructing our SVMs, we atempt 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.

The overall result of this study is widely open for improvements. More data in Indonesian can be obtained through the same approach or else in answering the scarcity of resource, providing more data to train and develop the emotion recognizer with. Furthermore, a more complex emotion definition 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.

Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 26870371.

7. REFERENCES

[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] Dessi Puji Lestari, Koji Iwano, and Sadaoki Furui, "A large vocabulary continuous speech recognition system for Indonesian language," in *15th Indonesian Scientific Conference In Japan*, 2006, pp. 17–22.

[6] Sakriani Sakti, Arry Akhmad Arman, Satoshi Nakamura, and Paulus Hutagaol, "Indonesian speech recognition for hearing and speaking impaired people.," in *INTERSPEECH*, 2004.

[7] S Sakti, M Paul, R Maia, S Sakai, N Kimura, Y Ashikari, E Sumita, and S Nakamura, "Toward translating Indonesian spoken utterances to/from other languages," in *Proceedings of O-COCOSDA*, 2009, pp. 137–142.

[8] Hammam R Yusuf, "An analysis of indonesian language for interlingual machine-translation system," in *Proceedings of the 14th conference on Computational linguistics-Volume 4*. Association for Computational Linguistics, 1992, pp. 1228–1232.

[9] S Sakti, R Maia, S Sakai, T Shimizu, and S Nakamura, "HMM-based speech synthesis of indonesian language," in *Proceedings of ASJ Spring Meeting*, 2008, pp. 301–302.

[10] 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.

[11] Bjoern Schuller, Gerhard Rigoll, and Manfred Lang, "Hidden Markov model-based speech emotion recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 2, pp. II–1.

[12] Laurence Devillers, Lori Lamel, and Ioana Vasilescu, "Emotion detection in task-oriented spoken dialogues," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*. IEEE, 2003, vol. 3, pp. III–549.

[13] Tim Polzehl, Alexander Schmitt, and Florian Metze, "Approaching multi-lingual emotion recognition from speech-on language dependency of acoustic/prosodic features for anger recognition," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[14] 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.

[15] 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.

[16] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 865–868.

[17] 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.

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

[19] 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.

[20] 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.