IEICE TRANS. INF. & SYST., VOL.E101–D, NO.8 AUGUST 2018

PAPER

Construction of Spontaneous Emotion Corpus from Indonesian TV Talk Shows and Its Application on Multimodal Emotion Recognition

Nurul LUBIS[†a)], **Dessi LESTARI**[††], ***Nonmembers***, **Sakriani SAKTI**[†,†††], ***Member***, **Ayu PURWARIANTI**[††], ***Nonmember***, ***and*** **Satoshi NAKAMURA**[†,†††], ***Member***

SUMMARY As interaction between human and computer continues to develop to the most natural form possible, it becomes increasingly urgent to incorporate emotion in the equation. This paper describes a step toward extending the research on emotion recognition to Indonesian. The field continues to develop, yet exploration of the subject in Indonesian is still lacking. In particular, this paper highlights two contributions: (1) the construction of the first emotional audio-visual database in Indonesian, and (2) the first multimodal emotion recognizer in Indonesian, built from the aforementioned corpus. In constructing the corpus, we aim at natural emotions that are corresponding to real life occurrences. However, the collection of emotional corpora is notably labor intensive and expensive. To diminish the cost, we collect the emotional data from television programs recordings, eliminating the need of an elaborate recording set up and experienced participants. In particular, we choose television talk shows due to its natural conversational content, yielding spontaneous emotion occurrences. To cover a broad range of emotions, we collected three episodes in different genres: politics, humanity, and entertainment. In this paper, we report points of analysis of the data and annotations. The acquisition of the emotion corpus serves as a foundation in further research on emotion. Subsequently, in the experiment, we employ the support vector machine (SVM) algorithm to model the emotions in the collected data. We perform multimodal emotion recognition utilizing the predictions of three modalities: acoustic, semantic, and visual. When compared to the unimodal result, in the multimodal feature combination, we attain identical accuracy for the arousal at 92.6%, and a significant improvement for the valence classification task at 93.8%. We hope to continue this work and move towards a finer-grain, more precise quantification of emotion.

key words: *Indonesian, audio-visual corpus, multimodal emotion recognition, spoken language processing*

1. Introduction

Emotion is an aspect yet to be fully replicated that is able to provide a richer and more natural human-computer interaction. Particularly in recent decades, the field has witnessed substantial advancements of studies and research on emotion. Further, researchers and scholars are putting the effort into real life applications through the construction of complex and emotionally advanced system, in particular spoken dialogue systems, e.g. Sensitive Artificial Listener [1], personable in-car assistant [2], and an embodied conversational companion [3].

This increasing interest in the topic is partly owing to the challenges held globally. In 2009 and 2010, INTERSPEECH spotlighted research on emotion in their annual challenges [4], [5]. With stricter focus on emotion recognition, the Audio Visual Emotion Challenge (AVEC) is held annually to address issues and unify solutions in the field. It is held for the first time in 2011, aiming at multimedia processing and machine learning methods for automatic emotion analysis [6]. The challenge developed from discrete emotional states into real time dynamic values in 2012 [7], up until the inclusion of physiological signals in 2015 and 2016 [8], [9].

However, the result of these advancements, challenges, and research on emotion are often language dependent. Unfortunately, adaptation of these works into a new language is not straightforward; a study of English-German multilingual emotion recognition using interactive voice response (IVR) recordings argues that for different languages, different sets of features are needed to obtain optimal emotion recognition performance [10]. Even though efforts on cross-lingual recognition is underway, uniformly high performance across the languages is still difficult to obtain [11]. This means, language specific research is necessary for application in a certain language.

In Asian languages, findings in affective computing continue to emerge. In Chinese, speech based emotion recognition for three emotion classes has been researched [12]. In Tagalog, an automated narrative storyteller was constructed with avarage precision of 86.75% in expressing a particular emotion [13]. On the other hand, in Indonesian, even though spoken language technology starts to actively develop since the past few years, research on emotion recognition is still non-existent. Currently, there exist a number of studies on automatic speech recognition [14], machine translation [15], and speech synthesis [16]. Unfortunately, for emotion recognition, even the resource to conduct studies and research on is still very lacking.

This is the reason we initiate the research on emotion recognition in Indonesian. In this effort, emotional corpora act as the starting point. As most approaches are supervised,

Manuscript received November 6, 2017.
Manuscript revised March 26, 2018.
Manuscript publicized May 10, 2018.

[†]The authors are with the Augmented Human Communication Lab, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
[††]The authors are with the School of Informatics and Electrical Engineering, Institut Teknologi Bandung, Indonesia.
[†††]The authors are with the RIKEN, Center for Advanced Intelligence Project AIP, Ikoma-shi, 630–0192 Japan.
a) E-mail: nurul.lubis.na4@is.naist.jp

DOI: 10.1587/transinf.2017EDP7362

the necessity of training data for experiments and modeling is non-negotiable. Moreover, the nature of the data will determine the quality of the system. This means, it is important to collect natural data that has as small a gap as possible with real life emotion occurrences.

However, the collection of such emotional data for research purposes is a sensitive matter, often raising moral and ethical issues [17]. Spontaneous emotion occurrence is likely to be kept private, as it is inherently a personal human experience. On the other hand fabrication of emotion is non-trivial and it is extremely difficult to match real life emotion. To tackle this, we construct the corpus in Indonesian from TV programs, particularly talk shows, containing real conversations and spontaneous emotions. The format of TV talk shows represents a typical social conversation between a small group of people, where various social topics that elicit story-telling, opinions, and emotions are discussed.

In the experiment, we utilize the collected data and employ the support vector machine (SVM) algorithm to model the emotions contained within. We perform multimodal emotion recognition utilizing the predictions of three modalities: acoustic, semantic, and visual. In this paper, we report the result of our corpus construction and emotion recognition experiment, followed with detailed analyses.

2. Related Work

Emotions in Indonesian language have been previously studied from the point of view of psychology [18], [19]. In affective computing, a number of works in Indonesian have not yet emerged until recently. One of the latest works on Indonesian Twitter data proposed an emotion lexicon argued to boost sentiment analysis performance [20]. Similarly, [21] attempted to recognize emotion from Indonesian tweets using linguistic, semantic, and orthographic features. On the other hand, [22] proposed a gamelan (Indonesian traditional instrumental orchestra) music emotion recognition for a robot puppeteer, able to distinguish delightful song, fearful song, and noise.

However, to our knowledge, the existing works in Indonesian are limited to unimodal recognition on textual and musical data. As studies on emotion recognition ultimately aims at enhancing the quality of human-computer interaction (HCI), recognition on fundamental communication channels such as acoustic and visual is essential to accommodate natural spoken language interaction between a user and a system [23], [24].

Toward this goal, we extend the works in Indonesian for spoken language by utlizing multimodal features. In this work, we first construct an Indonesian multimodal corpus suitable for real life application. We adapt the approach in [25] and utilize television program recordings, to construct the first audio-visual emotional corpus in Indonesian. We subsequently conduct experiments with the constructed corpus using SVM by utilizing three different facets of information, potentially containing emotional clues: acoustic, semantic, and visual. The combination of these modalities is hoped to provide a fuller observation of emotion occurrences.

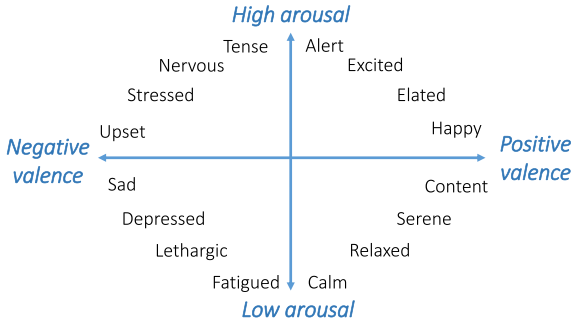

Fig. 1 Emotion dimensions and common terms.

3. Emotion Definition

Defining and structuring emotion is essential in observing and analyzing its occurrence. In this work, we define the emotion scope based on the circumplex model of affect [26]. This model highlights the emotion felt as opposed to underlying process of it, and is able to represent both primary and secondary emotion. An advantage of this model is that it is intuitive and also easily adaptable and extendable to either discrete or dimensional emotion definitions. The long established dimension are core to many works in affective computing and potentially provides useful information even at an early stage of research. Furthermore, this allows useful comparison to a wide range of related works.

Two dimensions of emotion are defined: valence and arousal. Valence measures positivity or negativity of emotion; e.g. joy is indicated by positive valence while fear is negative. On the other hand, arousal measures the activity of emotion; e.g. depression is low in arousal (passive), while rage is high (active). Figure 1 illustrates the valence-arousal dimension in respect to a number of common emotion terms.

4. Construction of Indonesian Spontaneous Emotional Audio-Visual Corpus

Although there has been an increase of interest in constructing corpora containing social interactions [27], [28], there is still a lack of spontaneous and emotionally rich corpora especially in low-resource languages. To bridge this gap, we construct a corpus of spontaneous social-affective interaction in the wild. We utilize various television talk shows containing natural conversations and real emotion occurrences. The steps of corpus construction are visualized in Fig. 2, and each will be described in detail in this section.

4.1 Spontaneous Emotional Data Collection

Many of emotion-related works rely on acted emotional data. While this provides rich emotional content, the portrayed emotions differ considerably from real occurrences in

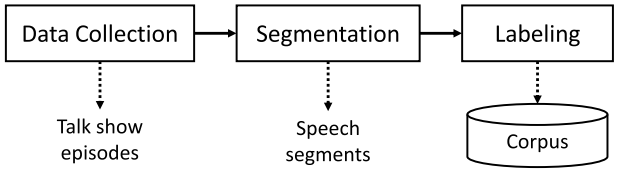

Fig. 2 The steps of corpus construction

everyday life, causing a potential mismatch when the technology is applied to real human-computer interaction. In this paper, we utilize emotion occurrences that is closer to the natural ones by utilizing television talk show recordings.

We collect the spontaneous emotional data in Indonesian from various television programs, in particular talk shows. This approach is previously utilized on the "Vera am Mittag" corpus [25]. Interactions in talk show setting well represent typical social conversation, where a small number of speakers are involved and various emotion-inducing topics are discussed. Furthermore, television talk shows provide clean recordings with distinguishable dialogue turns, resulting in good quality conversational data. As interactions are not scripted, the format gives us natural utterances with spontaneous emotion occurrences. Television talk shows also provide speaker variance. With different guests in each episode, we are able to gather data from a variety of speakers.

We select three episodes from different kinds of talk shows to cover a broader range of emotion. The selected talk shows are well-known in the country with high viewership, containing discussions on engaging and interesting topics that trigger various emotions from the speakers. The first show is "Mata Najwa," with discussion 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 lifestyle. The different topics are expected to provide more varied emotional content in the collected data.

The three talk show episodes are 2 hours, 25 minutes, and 39 seconds in length. There are 18 speakers in total; 12 male and 6 female. Even though the speech-speaker ratio is not uniform due to the different speaker roles in the talk show, this nonetheless offers speaker variance in the corpus.

4.2 Annotation

4.2.1 Labels

We perform the annotation on the sentence level in terms of emotional states. We define a sentence as a continuous utterance of a speaker until either one of the following conditions is met: 1) a full sentence is produced, or 2) it is proceeded by a different speaker. Segmentation with these criteria results in 847 sentences, or speech segments, in the corpus.

As elaborated in Sect. 3, in this paper we follow the circumplex model of affect as the computational model of emotion [26]. We define two emotion dimensions as the descriptor of felt emotion; valence and arousal. Valence measures the positivity or negativity of emotion while arousal measures the activity. Following this model, the emotion annotation of the corpus consists of the level of valence (`val`) and (`aro`). The value of each dimension can be as low as −3 and as high as 3 with a discrete step of 1. This provides a granularity that balances between details of information and cognition load for the annotators.

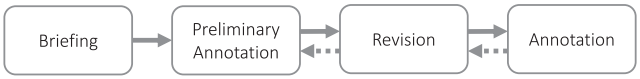

Fig. 3 Overview of annotation procedure

4.2.2 Procedure

In annotating the corpus, we bear in mind that language and culture affect how emotion is perceived and expressed in an interaction. We carefully select 3 human annotators. Every annotator is required to be (1) a native Indonesian speaker, and (2) knowledgeable of the Indonesian culture of communication. With these requirements, we try to ensure that the annotators can observe emotion dynamics of the interaction to the furthest extent in addition to recognizing the emotion appropriately. To ensure consistency, we have each annotator annotate the full corpus.

Figure 3 gives an overview of the annotation procedure. Before annotating the corpus, the annotators are briefed and given a document of guidelines to get a clearer picture of the task and its goal. The document provides theoretical background of emotion in discourse as well as a number of examples.

After the annotators are briefed, firstly, we ask them to do preliminary annotation by working on a small subset of the corpus. This step is done to let them get familiar with the task. Furthermore, with the preliminary result, we are able to confirm whether the annotators have fully understood the guidelines, and verify the quality and consistency of their annotations.

We manually screen the preliminary annotation result and give feedback to the annotators accordingly. They are asked to revise inconsistencies with the guidelines if there are any. This process is repeated until the quality of the preliminary annotation is sufficient. Once their results are verified, the annotators are authorized to work on the rest of the corpus. We perform the same screen-and-revise process on the full corpus annotation to achieve a tenable result.

5. Corpus Analysis

We inspect properties of the corpus to gain better insight of the data contained within. We look over the conversational aspect of the collected data through the distribution of segments length in the corpus. Furthermore, we examine the quality of emotion annotations by looking at the inter-annotator agreement.

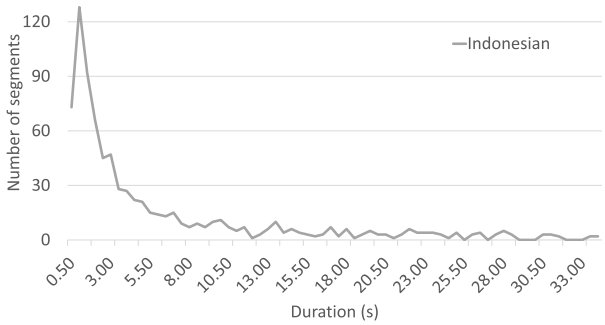

Fig. 4 Number of segments in respect to the duration

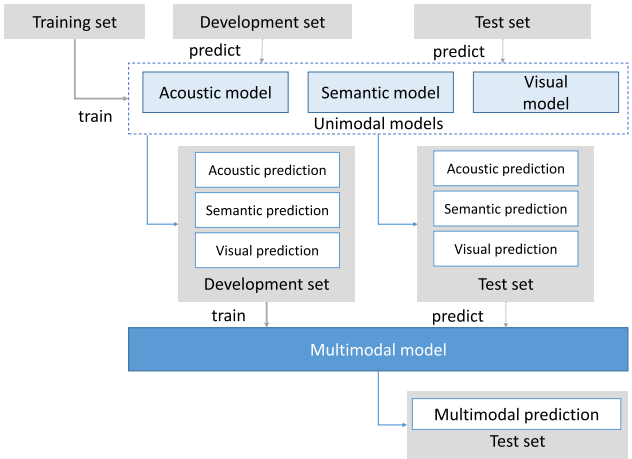

Fig. 5 Combination scheme of multiple modalities

5.1 Length of Segments

We plot the distribution of number of segments according to their duration on Fig. 4. In the figure, the y-axis shows the number of segments with respect to the x-axis, which shows the duration. The segments have high variance of duration with a peak at around 1 second. The average duration of segments is 7.57 seconds.

5.2 Annotators Agreement

To analyze the consistency of annotation, we investigate the annotators agreement for both emotion dimension annotations. We calculate mean Pearson's correlation coefficients r of the three annotators for each emotion dimension. Pearson's r measures the strength and direction of linear relationship between two variables. An absolute value of r between 0.0 and 0.3 is interpreted as weak correlation, and greater than 0.3 up to 0.5 as moderate correlation. For `val` and `aro`, we achieve moderate correlations of 0.33 and 0.37, respectively. This shows similar levels of agreement between the annotators in perceiving both valence and arousal, suggesting uniform capabilities in perceiving both dimension in the sentence segments.

5.3 Emotion Label Distribution

We average the annotation of the 3 annotators to obtain the average valence and arousal of a segment. The majority of the collected segments lies in the $[0, 1]$ (43.80% for valence and 46.22% for arousal) and $[1, 2]$ (31.28% for valence and 45.33% for arousal) value ranges. There is little coverage of the extreme values (-3 and 3) and the negative spectrum of either dimensions. This observation shows some bias towards the positive regions for both valence and arousal.

A possible explanation for this is the nature of television talk shows in the first place. Discussions in the television talk shows are normally designed to be engaging and enjoyable for the viewers, as is the case with the collected episodes. In such situations, the occurrence of deactivated (negative arousal) and negative (negative valence) emotions appear to be rare. This points to the limitation of the conversational scope of television talk shows, which should be addressed upon corpus expansion in the future.

6. Multimodal Emotion Recognition

We attempt to train a high-accuracy emotion recognizer by performing a feature combination of different modalities. In the following experiments, we train unimodal classifiers using acoustic, semantic, and visual features, and fuse them into a multimodal one by combining their predictions.

Figure 5 visualizes the combination scheme proposed in this work. The model construction starts with the training of several classifiers using various unimodal features. We then utilize these unimodal models to make prediction on subsets of the corpus. We utilize their predictions, in the form of probability distributions, as classification features for the multimodal model. By including features of different modalities, we hope to be able to observe a more complete picture of emotion occurrences. To the best of our knowledge, this constitutes the first multimodal emotion recognition effort in Indonesian.

We perform feature combination at decision level as described above, instead of at feature level by concatenating the features, due to the small amount of data available. Feature-level fusion increases the number of features during training, even though the amount of data remains the same. Furthermore, any modification of features will lead to retraining of the entire model. On the other hand, the decision level fusion treats the modalities in a more modular way, allowing modification without the need of full model retraining.

The use of semantic or lexical information for emotion recognition commonly requires a large, often knowledge-based, language-specific database, i.e. word2vec [29] or WordNet Affect [30]. The construction of such model can be problematic for many languages where resources are scarce, such as Indonesian. In this work, in addition to the acoustic and visual features, we utilize semantic information ex-

tracted purely from the available emotion corpus, without the need of additional data. We exploit the TF-IDF score to weight the words to form the semantic representation.

7. Experiments

7.1 Set-Up

In this experiment, we utlize the TV talk show corpus previously constructed to train the emotion recognizer. Due to the relatively small amount of data, we simplify the emotion recognition problem by discretizing the affective dimensions values into three classes: positive, neutral, and negative. This simplification is a trade-off decision between recognition granularity and model complexity. This discretization scheme has also been used in a number of previous studies, such as [31]. We believe that incremental steps in tackling the problem will result in a better recognizer in the long run, thus we focus on the less complex problem at this stage of the research as the foundation for future experiments.

The emotion label of each segment is decided by majority voting. We first discretize the annotation into positive, neutral, and negative, and perform majority voting afterwards. The positive class corresponds to the positive-valued annotation (3 to 1), the neutral class to the zero-valued (0), and negative class to the negative-valued (-1 to -3). Segments with three different votes are excluded to avoid potentially ambiguous emotion occurrences. We obtain 805 (555 positive, 119 neutral, and 131 negative) segments for valence and 820 (691 positive, 128 neutral, and 1 negative) for arousal. We randomly divide the total with 80:10:10 ratio into training, development, and test sets.

We extract several feature sets to represent three fundamental modalities of communication: acoustic, semantic, and visual. First, we extract global features of each utterance using the openSMILE toolkit [32]. Two different acoustic feature sets are selected: INTERSPEECH 2009 baseline features (IS09) and extended Geneva Minimalistic Acoustic Parameter Set (eGemaps) feature sets. The IS09 feature set is described in Table 1. This feature set is widely used in emotion recognition research, thus providing comparability to extensive related works.

On the other hand, the eGemaps feature set is proposed as reduced acoustic feature set, containing only knowledge-based selected features that are 1) highly potential in indexing affective signals, and 2) proven to be effective in previous studies [33]. This feature set includes parameters related to frequency (pitch, jitter, formant), energy (shimmer, loudness, HNR), spectral (alpha ratio, Hammarberg index, MFCC 1-4, spectral scope, format relative energy and bandwidth, spectral flux). Reduction of the total numbers of features is mostly contributed by the much fewer number of functionals, mostly only involving arithmetic mean and coefficient variation, with additional ones for selected features.

Secondly, to extract the semantic features of the utterances, we compute the TFIDF weighted vector of its transcribed speech. The TFIDF weight of a term t in a sentence T is computed as:

$$\mathrm{TFIDF}(t, T) = F_{t,T} \log \frac{|T|}{DF_t}, \tag{1}$$

where $F_{t,T}$ is defined as term frequency of term t in a sentence T, and DF_t as total number of sentences that contains the term t, calculated over the training set. Thus, the vector for each sentence is the size of the corpus term vocabulary, with each term weighted according to Eq. (1).

Lastly, we extract facial features of each frame with the OpenFace toolkit [34]. We extract 2D position of the facial landmarks, as well as Action Unit (AU) intensities, and treat them as two separate feature sets. We exclude frames with low face detection confidence to exclude frames containing other objects and uncertain recognition results. To obtain the segment level features, we calculate 4 statistics from the collection of frames for a given speech segment: mean, standard deviation, minimum, and maximum values. Since the frames are collected at rapid intervals, this helps neutralize the effect of multiple faces in a segment should it occur. Table 2 summarizes the total number of features in each feature set.

We train an SVM classifier for each feature set and combination of the three modalities. We use the libSVM library [35] in both of our unimodal and multimodal experiments. Prior to training, we scale the features into $\{0, 1\}$ range to avoid overpowering of features ranging in big values. Furthermore, we perform parameter optimization with grid search to find the optimal value of C, the cost of misclassification, and γ, the free parameter of the Gaussian RBF kernel. These steps are recommended in [36] and has been shown to be effective in SVM classification experiments. For all models, unimodal or multimodal, we compare their performance by using the recognition precision, recall, and F1-score on the test set.

7.2 Results and Discussion

Table 3 presents the performance of unimodal emotion recognition, measured in precision, recall, and F1 score. We

Table 1 Baseline feature of INTERSPEECH 2009 emotion challenge

LLD ($16 \cdot 2$)	Functionals (12)
($\triangle$) ZCR	mean
($\triangle$) RMS Energy	standard deviation
($\triangle$) F0	kurtosis, skewness
($\triangle$) HNR	extremes: value, rel. position, range
($\triangle$) MFCC 1-12	linear regression: offset, slope, MSE

Table 2 Number of features in each feature set

Modality	Feature set	Number of features
Acoustic	gemaps	62
	IS09	384
Visual	AU	72
	landmarks	336
Semantic	TFIDF	2801

Table 3 Unimodal emotion recognition performance on test set in %. Highest number on each task is boldfaced.

Modality	Feature	Arousal			Valence		
		Prec.	Rec.	F1	Prec.	Rec.	F1
Acoustic	eGemaps	90.02	90.24	89.12	87.23	83.95	85.19
	IS09	93.28	92.68	**91.84**	90.23	86.42	**87.87**
Visual	AU	68.77	82.93	75.19	66.39	81.48	73.17
	landmarks	68.77	82.93	75.19	66.39	81.48	73.17
Semantic	TFIDF	68.77	82.93	75.19	79.69	53.09	60.97

Table 4 Multimodal emotion recognition performance on test set in %. Highest number on each task is boldfaced. *: higher than unimodal best.

Features			Arousal			Valence		
Semantic	Acoustic	Visual	Prec.	Rec.	F1	Prec.	Rec.	F1
TFIDF	eGemaps	AU	68.76	82.92	75.18	84.12	83.95	78.45
		landmarks	68.76	82.92	75.18	84.12	83.95	78.45
	IS09	AU	93.27	92.68	**91.83**	91.79	93.82	**92.46**
		landmarks	93.27	92.68	**91.83**	90.25	92.59	91.26

average the measurements of the three classes with weighted averaging based on class size to take into account the class imbalance. The unimodal recognition result shows that in both tasks acoustic signals appear to be the most discriminative classification feature, suggesting that the richest emotion clue in the conversational data is carried through speech. Interestingly, regardless of the much bigger number of features, the IS09 set outperforms the eGemaps set. This suggest that some features that are present in IS09 but absent in eGemaps are helpful in recognizing emotion in the constructed Indonesian corpus.

For both tasks, the recognition performance using visual and TFIDF features is slightly lower than that of acoustic. The two visual feature sets yield identical performance, with precision score that are considerably lower than the recall. Reversely, the TFIDF feature set scores higher in precision than recall. This shows an imbalance in the prediction that favors or misses the larger class. On the other hand, the acoustic features achieve high score for both recall and precision.

Table 4 summarizes the result of multimodal emotion recognition. We experimented with all possible combinations of the three modalities. The numbers suggest that the determining factor in the result of feature combination is the highest performing feature among the unimodal ones. For example, the decision of which visual features to use in the combination does not affect the performance of the end multimodal model, most probably due to its suboptimal performance in comparison to the other unimodal features.

Due to the lack of negative sample on the collected data, the arousal models essentially learned only two classes (positive and neutral). In the arousal task the best combination of the modalities yields identical number as that of the unimodal. On the other hand, in the valence task, the combination is able to improve the recognition rate significantly. To understand this difference better, we analyze the result of the unimodal and multimodal recognition. For each pair of the best performing multimodal sets (i.e. set ABC becomes pairs AB, AC, and BC), we compute the correlation coefficient of the probability distributions to see the similarity of the predictions. We find that on the arousal task, the unimodal predictions are highly correlated, making most mistakes on the same examples. In consequence, they do not provide additional information to each other when combined. In contrast, for the valence task, we observe variance and different trends of prediction. This means the modalities contain complimentary information that is useful in better modeling the problem and classifying the data on the second stage of the model training.

The improvement of emotion recognition performance after considering multimodal information has been shown in previous studies in other languages as well. In a study using an acted emotion corpus, Busso et al. reported an accuracy of 89% in recognizing four classes of emotion, a significant improvement compared to the accuracy of 70.9% with speech-based system and 85% with facial expression based system [37]. This finding is reaffirmed by Poria et al. who reported improvement when fusing three modalities into a single emotion recognizer [38]. Another study by Nojavanasghari et al. focusing on spontaneous emotions in children also reported the same trend, with best overall performance at 69% [39].

However, comparison between emotion recognition results should be done carefully. Comparability has been a long standing issue for emotion recognition tasks. The different methods and approaches in constructing an emotion recognizer is difficult to compare as they are rarely tested on a common experimental condition [40]. One of the reason of this problem is the absence of a database that can generalize to all facets of human emotion (e.g. culture, gender, situation) to act as the standard corpus of emotion recognition works [41]. Each emotion-rich corpus is often aimed to capture emotion in a specific set of circumstances, creating not only differences between the corpora, but also the works based on them. Furthermore, different emotion models result in a variety of annotation scheme (e.g. dimensional traces or emotion classes) and recognition problems (e.g. classification or regression). Readers are referred to litera-

tures for comprehensive survey and comparison on emotion recognition works [17], [41].

8. Conclusion and Future Works

We presented an emotional audio-visual corpus in Indonesian and the subsequent effort utilizing it in multimodal emotion recognition. To construct the corpus, we exploited television programs, in particular talk shows, to provide natural emotional data covering a broad range of emotions. The corpus is the first of its kind in Indonesian, providing a wide-range of opportunities of studies on emotion in the language, especially as very few resources currently exist. Subsequently, we construct an SVM based multimodal emotion recognizers utilizing the predictions of three modalities: acoustic, semantic, and visual. We perfom decision-level combination and obtained a model with identical or better recognition accuracy.

In this paper we have successfully utilized a corpus constructed from spontaneous conversation to build emotion recognizers for three levels of valence and arousal. We hope to continue this work and move towards a finer-grain, more precise quantification of emotion. Continuous development of the corpus is important, and we hope to collect more data from other situated dialogue to cover a wider scope emotion and of social interactions. In the future, the content of the conversation should be considered during data collection to address the uneven distribution of emotion labels we have reported in this paper. We look forward to experiment with other features and methods to improve on the recognition accuracy.

Acknowledgements

Part of this work was supported by JSPS KAKAENHI Grant Numbers JP 17H06101 and JP 17K00237.

References

[1] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M.T. Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wollmer, "Building autonomous sensitive artificial listeners," IEEE Trans. Affective Computing, vol.3, no.2, pp.165–183, 2012.

[2] K.J. Williams, J.C. Peters, and C.L. Breazeal, "Towards leveraging the driver's mobile device for an intelligent, sociable in-car robotic assistant," 2013 IEEE Intelligent Vehicles Symposium (IV), pp.369–376, IEEE, 2013.

[3] D. Benyon, B. Gambäck, P. Hansen, O. Mival, and N. Webb, "How was your day? evaluating a conversational companion," IEEE Trans. Affective Computing, vol.4, no.3, pp.299–311, 2013.

[4] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," INTERSPEECH, pp.312–315, Citeseer, 2009.

[5] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," INTERSPEECH, pp.2794–2797, 2010.

[6] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," Affective Computing and Intelligent Interaction, pp.415–424, Springer, 2011.

[7] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," Proc. 14th ACM international conference on Multimodal interaction, pp.449–456, ACM, 2012.

[8] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," Proc. 5th International Workshop on Audio/Visual Emotion Challenge, pp.3–8, ACM, 2015.

[9] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M.T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," AVEC '16 Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp.3–10, ACM 2016.

[10] T. Polzehl, A. Schmitt, and F. Metze, "Approaching multi-lingual emotion recognition from speech-on language dependency of acoustic/prosodic features for anger detection," 2010.

[11] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," 2016 IEEE International Conference on Acoustics, Speech Signal Process. (ICASSP), pp.5800–5804, IEEE, 2016.

[12] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," Int. J. Smart Home, vol.6, no.2, pp.101–108, 2012.

[13] J.C.P. Gonzaga, J.A. Seguerra, J.A. Turingan, M.P.A. Ulit, and R.A. Sagum, "Emotional techy basyang: An automated filipino narrative storyteller," Int. J. Future Computer and Communication, vol.3, pp.271–274, Aug. 2014.

[14] S. Sakti, A.A. Arman, S. Nakamura, and P. Hutagaol, "Indonesian speech recognition for hearing and speaking impaired people," INTERSPEECH, 2004.

[15] 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," Proc. O-COCOSDA, pp.137–142, 2009.

[16] E. Cahyaningtyas and D. Arifianto, "HMM-based indonesian speech synthesis system with declarative and question sentences intonation," 2015 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pp.153–158, IEEE, 2015.

[17] M. El Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol.44, no.3, pp.572–587, 2011.

[18] P.R. Shaver, U. Murdaya, and R.C. Fraley, "Structure of the indonesian emotion lexicon," Asian journal of social psychology, vol.4, no.3, pp.201–224, 2001.

[19] K.R. Scherer, R. Banse, and H.G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," J. Cross-cultural psychology, vol.32, no.1, pp.76–92, 2001.

[20] F. Koto and M. Adriani, "Hbe: Hashtag-based emotion lexicons for twitter sentiment analysis," Proc. 7th Forum for Information Retrieval Evaluation, pp.31–34, ACM, 2015.

[21] J.E. The, A.F. Wicaksono, and M. Adriani, "A two-stage emotion detection on indonesian tweets," 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp.143–146, IEEE, 2015.

[22] T.P. Tomo, G. Enriquez, and S. Hashimoto, "Indonesian puppet theater robot with gamelan music emotion recognition," 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp.1177–1182, IEEE, 2015.

[23] L. Devillers, M. Tahon, M.A. Sehili, and A. Delaborde, "Inference of human beings emotional states from speech in human–robot interactions," Int. J. Social Robotics, vol.7, no.4, pp.451–463, 2015.

[24] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. Heylen, "The sensitive artificial listener: an induction technique for generating emotionally coloured conversation," 2008.

[25] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," Proc. IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2008.
[26] J.A. Russell, "A circumplex model of affect," J. Personality and Social Psychology, vol.39, no.6, p.1161, 1980.
[27] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," Affective computing and intelligent interaction, pp.488–500, Springer, 2007.
[28] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp.1–8, IEEE, 2013.
[29] B. Xue, C. Fu, and Z. Shaobin, "A study on sentiment computing and classification of sina weibo with word2vec," 2014 IEEE International Congress on Big Data, pp.358–363, IEEE, 2014.
[30] C. Strapparava and A. Valitutti, "Wordnet affect: An affective extension of wordnet," LREC, pp.1083–1086, 2004.
[31] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," 2016 IEEE Spoken Language Technology Workshop (SLT), pp.565–572, IEEE, 2016.
[32] F. Eyben, M. Woellmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," Proc. international conference on Multimedia, pp.1459–1462, ACM, 2010.
[33] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, and K.P. Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," IEEE Trans. Affective Computing, vol.7, no.2, pp.190–202, 2016.
[34] T. Baltrušaitis, P. Robinson, and L.P. Morency, "Openface: An open source facial behavior analysis toolkit," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp.1–10, IEEE, 2016.
[35] C.C. Chang and C.J. Lin, "LibSVM: A library for support vector machines," ACM Trans. Intelligent Systems and Technology (TIST), vol.2, no.3, p.27, 2011.
[36] C.W. Hsu, C.C. Chang, and C.J. Lin, "A practical guide to support vector classification," 2003.
[37] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," Proc. 6th international conference on Multimodal interfaces, pp.205–211, ACM, 2004.
[38] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," 2016 IEEE 16th International Conference on Data Mining (ICDM), pp.439–448, IEEE, 2016.
[39] B. Nojavanasghari, T. Baltrušaitis, C.E. Hughes, and L.P. Morency, "Emoreact: A multimodal approach and dataset for recognizing emotional responses in children," Proc. 18th ACM International Conference on Multimodal Interaction, pp.137–144, ACM, 2016.
[40] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.1, pp.39–58, 2009.
[41] C.N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," Artificial Intelligence Review, vol.43, no.2, pp.155–177, 2015.

Nurul Lubis received the B.E degree (with distinction) in 2014 from Bandung Insistute of Technology, Indonesia and the M.Eng degree in 2017 from Nara Institute of Science and Technology (NAIST), Japan. She is currently a doctoral candidate at Augmented Human Communication Laboratory, NAIST, Japan. She is a recipient of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) scholarship. She was a research intern at Honda Research Institute Japan, Co. Ltd. Her research interest include affective computing, emotion in spoken language, and affective dialogue systems.

Dessi Lestari was born in Bandung, Indonesia, in 1979. She received the B.E. degree in informatics engineering from Bandung Institute of Technology (ITB), Indonesia, in 2003, and the M.Eng. and Ph.D. degrees in computer science from the Tokyo Institute of Technology (Titech) Tokyo, Japan, in 2007 and 2011, respectively. In 2012, she joined the Department of Informatics Engineering, School of Electrical and Informatics Engineering, ITB, as a Lecturer. Since February 2016, she became an Associate Professor in ITB. Her current research interests include speech signal processing, speech recognition, speech synthesizer, speaker recognition, emotional recognition, and broad domain of human computer interaction.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003–2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006–2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005–2008) with Dialog Systems Group University of Ulm, Germany, and received her Ph.D. degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003–2007), A-STAR and U-STAR (2006–2011). In 2009–2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. From 2011, she has been an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015–2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation". She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Ayu Purwarianti gained her bachelor and master degree at Informatics/Computer Science Program, Bandung Institute of Technology. She got her doctoral degree from Toyohashi University of Technology, Japan. Since 2008, she has become a lecturer at School of Electrical Enginering and Informatics, Bandung Institute of Technology, Indonesia. Now, she is the Chair of Doctoral Program of Electrical Engineering and Informatics at Bandung Institute of Technology, Indonesia. Her research interest is on computational linguistics, mainly on Indonesian natural language processing. She is now active as education officer at IEEE Indonesia and she is also active at Indonesian Association for Computational Linguistics.

Satoshi Nakamura is Professor of Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan, Project Leader of Tourism Information Analytics Team of RIKEN, Center for Advanced Intelligence Project AIP, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994–2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000–2008 and Vice president of ATR in 2007–2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009–2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampolli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.