CONSTRUCTION AND ANALYSIS OF INDONESIAN EMOTIONAL SPEECH CORPUS

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

In this paper we present Indonesian Emotional Speech Corpus (IDESC), the first ever corpus in Bahasa Indonesia that contains various emotion contents. As interaction between human and computer makes its way to the most natural form possible, it becomes more and more urgent to incorporate emotion in the equation. However, in Bahasa Indonesia, this aspect is yet to be explored. The acquisition of an emotion corpus serves as a foundation in further research regarding the subject. In constructing IDESC, we aim at natural and real emotion that is applicable to human-computer interaction. The corpus consists of three episodes of Indonesian talk show in different genres: politics, humanity, and entertainment. Each episode is carefully segmented and labeled based on its emotion content, resulting in 1357 segments worth 1 hour, 1 minute, and 43 seconds of speech. The corpus is still in its early stage of development, yielding exciting possibilities of future works.

Index Terms— Bahasa Indonesia, corpus, emotion, speech

1. INTRODUCTION

Emotion is an aspect yet to be fully replicated that is able to provide richer and more natural interaction between human and computer. Over the years, this issue continues to be addressed. This results in the development in the field of affective computing, through the construction of complex and emotionally advanced systems such as Sensitive Artificial Listener [1], personable in-car assistant [2], and even system that helps with emotional memory [3]. However, the majority of the advancements in affective computing are in English.

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 [4]. They continued to address affective issue in 2010 through one of their sub-challenges [5]. 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 [6]. After that, AVEC 2012 tried to analyze emotion from its dimensions rather than identifying it as discrete states [7].

In Asian languages, exists a number of studies and findings in affective computing. in Chinese, researchers have studied the effect of switching stimulus in user, involving affective system [8]. In Tagalog, an automated narrative storyteller was constructed with avarage precision of 86.75% in expressing a particular emotion [9]. Unfortunately, in Bahasa Indonesia, research on topics alike is almost non-existent– even the resource to conduct studies and research on is still very lacking.

This is the reason we initiate the construction of Indonesian corpus for human-computer interaction. However, data collection for this purpose is difficult as they have to mimic real emotion of system user. Most speech emotion corpus is collected through acting. While this provides prominent emotion content, the nature of the emotion does not match that of user's in their interaction with computer.

In details, we construct the speech corpus in Bahasa Indonesia from various talk shows, containing real conversations and real emotions. The construction aims at realistic emotion corpus that is applicable to human-computer interaction. We gather speech data from television broadcasts, segment and annotate them according to the emotion content. Each process in the construction of IDESC is done carefully manually through human recognition. After the construction, we perform analysis on the resulting corpus.

The remainder of this paper is organized as follows. Section 2 describes previous studies and works on emotion corpora. Section 3 explains the construction of IDESC. In section 4, we perform analysis on the constructed corpus. Section 5 concludes the paper with closing remarks and future work.

2. RELATED WORK

In terms of content, emotion corpora have been constructed from various sources and data collection methods. The employment of actors is popular in earlier construction to provide data with prominent emotion state. Researchers then shift to naturalistic data as it's potentially more relevant for affective computing.

The Danish Emotional Speech (DES) Database was collected from acted emotional utterances of actors and actresses [10]. Similarly, the Geneva Multimodal Emotion Portrayals (GEMEP) corpus was constructed from acted portrayals of emotion [11]. Acting and portrayals is a reasonable approach, especially in the case of affective corpus, as spontaneous emotion occurrences are hard to obtain. Some concerns are raised in the usage of acted affective corpus; it's stereotypical, lacking context, and limited to general and basic emotions. However, previous study had argued that if designed carefully, acted emotion portrayals can nonetheless be beneficial in emotion research [12].

Built with a different approach, The SEMAINE Database consists of natural dialogue between user and operator simulating a limited agent using Sensitive Artificial Listener (SAL) scenario [13]. Different SAL characters are present, eliciting different emotions in user, thus resulting in emotionally colorful corpus. This is one solution in providing naturalistic emotion content, by triggering certain emotion in participant. However, the set-up sometimes gives unpredictable result as the participants are fully aware of the intention of the agents.

Other non-acted emotion corpora include the Vera am Mittag corpus, collected from German television talk-show [14], and HUMAINE Database, a multimodal corpus consisting of naturalistic and induced data showing emotion in a range of contexts [15]. The usage of interview recordings is beneficial as they are natural and the emotion occurrences are more realistic. However, obtaining diverse emotion contents might be tricky as in such environment certain behaviors are undesirable.

The number of emotion corpora in oriental languages is not as many as that of English and European languages. Furthermore, the resource available is mostly limited to textual data. In Chinese, a textual emotion corpus was constructed [16]. This corpus contains manual annotation of eight emotion categories in intensity, target, word/phrase, and other linguistic expressions that indicate emotion. In Japanese, a speech corpus was constructed using online gaming voice chat, consisting both naturalistic and acted emotion [17].

The constructed corpus in this paper is the first emotional speech corpus in Bahasa Indonesia. It contains only naturalistic emotion occurrences, as we aim at applicability to human-computer interaction technologies.

3. IDESC: INDONESIAN EMOTIONAL SPEECH CORPUS

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 details in this section.

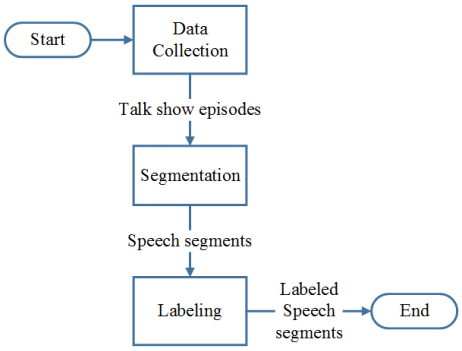

Fig. 1. The steps of corpus construction

3.1. Data Collection

We collect the speech data in Bahasa Indonesia from various television talk shows. This approach is done before on the "Vera am Mittag" corpus [14]. 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.

We select three episode from different kinds of talk shows to cover broader range of emotion. 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 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 life. The different topics are expected to provide more varied emotion content in the collected data.

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 is 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 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 is 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 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 references 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 label: 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 Bahasa Indonesia.

4. ANALYSIS

We perform statistical analysis on the length of the segments of speech. This may give us better insight of the nature of emotion occurrence in Bahasa Indonesia. The shortest segment is 0.16 second long while the longest is 15.47.The distribution of the length of speech is visualized in Fig. 2.

Most of the segments are short in length, with median of 2.20 seconds. The number of segments in a length interval reaches its peak between 1.00 to 1.49 seconds with total of 319 segments. After the peak, the number continues to decrease as the length increases. This statistic tells us that in speech, one certain emotion tend to occur only in a short period of time before any change or transition occurs, averaging at 2.60 seconds.

We also analyze IDESC on its emotion content. Firstly, we analyze the agreement level between human annotators using Fleiss' Kappa. These numbers are presented in Fig. 3.

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

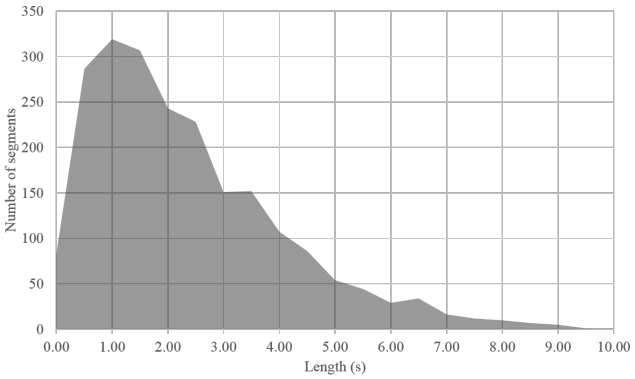

Fig. 2. Segment length distribution in the corpus

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.

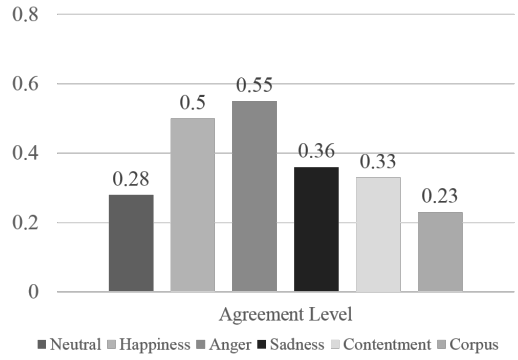

Fig. 3. Agreement level between annotators measured with Fleiss' Kappa

The overall agreement level of the annotators on the entire corpus reaches 0.23, which can be interpreted as slight agreement. We try to analyze this number further by looking at the label correctness of each annotator compared to the chosen labels. The numbers are presented in Fig. 4. This statistic confirms the variety of emotion perception by the annotators. This subjectivity of perception can be neutralized by employing more annotators for the corpus.

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. 5.

In general, neutral and passive-positive emotion seem to be the most common occurrence in the collected dialogues.

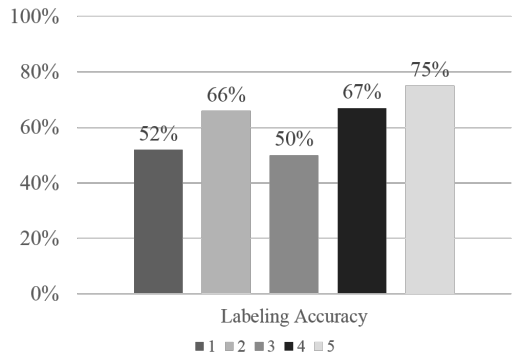

Fig. 4. Labeling accuracy of the annotators

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.

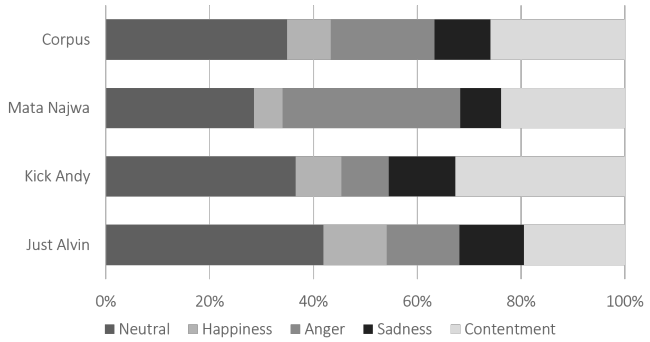

Fig. 5. Distribution of emotion labels in the corpus

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

Discussion wise, a causal relationship can be formed between the talk turns of the host and guest. A guest on the talk show will respond to the questions of the host, and the host will respond with more questions for the guest until eventually the topic shifted. This pattern occurs very often, thus IDESC may be beneficial for research regarding emotion triggers in dialogues. Table 1 contains transcription of dialogue example in the corpus that shows this pattern.

Host	Ayah Anda adalah pendidik, pejuang, ya? Diplomat juga. Tapi boleh tahu nggak, di mata Anda, seorang Sutan Rasyid ini, orang yang seperti apa sih? (*Your father is an educator, fighter, yes? And also a diplomat. But I'd like to know, in your eyes, who is your father, Sutan Rasyid? What kind of person is he?*)
Guest	Hmm kembali ke memori ya. Bung Andy, kalau saya coba ingat-ingat kembali, beliau itu orang yang keras. Seorang pendidik, karena saya tahu cerita beliau waktu kecil Ibunya itu sudah meninggal. Jadi diasuh oleh Tante. Jadi ceritanya Ayah kepada saya itu tidak pernah mendapat *mother's love*. (*Hmm, to turn back to memory, as I can recall, he is a tough man. An educator, cause from what I know from his stories, he lost his mother when he was little, and was raised by an aunt. So from his stories to me, he had never experienced mother's love.*)
Host	Baik. Dan Rasyid ini, karirnya dari apa dulu? (*I see. And him, Rasyid, how did his carreer begin?*)
Guest	Diplomat itu justru di masa yang terakhir itu, Bung Andy. Sebelumnya beliau itu adalah sekjen di Deplu. (*He wasn't a diplomat until the later days in his carreer, Andy. Before, he was a general secretary in the ministry of foreign affairs.*)

Table 1. A transcription of dialogue example from the corpus (in Bahasa Indonesia and English translation)

5. CONCLUSION AND FUTURE WORKS

This paper presents IDESC, the first corpus of emotional speech in Bahasa Indonesia. IDESC provides exciting opportunities in affective computing developments in Bahasa Indonesia, especially as very few researches currently exist. The analysis performed shows interesting tendencies of emotion occurrence in Indonesian dialogues and its correlation to the topic discussed.

In developing IDESC, we will collect more data to obtain more variety in speech and emotions. Furthermore, we will try to give more detailed annotation of the corpus, e.g transcription and quantification of the emotion occurrence. More human reference will be employed and simultaneously experiment on the usage of tools to automatize the process will be conducted. This way, more analyses and applications of IDESC will be made possible.

In the future, we will make use of IDESC to perform speech based automatic emotion recognition in Bahasa Indonesia. During the process, we will also analyze the acoustic features that are important in doing so, giving us more insight

on prosodic characteristics of Bahasa Indonesia. This will hopefully initiate further studies, research, and findings on affective computing in Bahasa Indonesia, which is what the constructed corpus aims at in the first place.

Acknowledgements

Part of this research is supported by JSPS KAKENHI Grant Number 26870371.

6. REFERENCES

[1] Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, et al., "Building autonomous sensitive artificial listeners," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 165–183, 2012.

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

[3] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski, "Affectaura: an intelligent system for emotional memory," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 849–858.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Inger S Engberg, Anya Varnich Hansen, Ove Andersen, and Paul Dalsgaard, "Design, recording and verification of a danish emotional speech database," in *Eurospeech*, 1997.

[11] Tanja Baenziger, Hannes Pirker, and K Scherer, "Gemep-geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions," in *Proceedings of LREC*, 2006, vol. 6, pp. 15–019.

[12] Tanja Baenziger and Klaus R Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus," in *Affective computing and intelligent interaction*, pp. 476–487. Springer, 2007.

[13] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroeder, "The semaine database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, pp. 5–17, March 2012.

[14] Michael Grimm, Kristian Kroschel, and Shrikanth Narayan, "The vera am mittag german audio-visual emotional speech database," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2008.

[15] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret McRorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, Noam Amir, and Kostas Karpouzis, "The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data," *Affective Computing and Intelligent Interaction*, pp. 488–500, 2007.

[16] Changqin Quan and Fuji Ren, "Construction of a blog emotion corpus for chinese emotional expression analysis," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 2009, pp. 1446–1454.

[17] Y Arimoto and H Kawatsu, "Online game emotional speech corpus using voice chat," in *Proceedings of 2013 Autumn Meeting Acoustical Society of Japan*, pp. 385–388.