Construction of English-French Multimodal Affective Conversational Corpus from TV Dramas

Sashi Novitasari[1,2], Quoc Truong Do[1], Sakriani Sakti[1], Dessi Lestari[2], and Satoshi Nakamura[1]
[1]Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, Japan
{do.truong.dj3,ssakti,s-nakamura}@is.naist.jp
[2]Department of Informatics, Bandung Institute of Technology, Indonesia
Jl. Ganesha No.10, Bandung, Jawa Barat, Indonesia
{sashinovitasari11,dessipuji}@gmail.com

Abstract

Recently, there has been an increase of interest in constructing corpora containing social-affective interactions. But the availability of multimodal, multilingual, and emotionally rich corpora remains limited. The tasks of recording and transcribing actual human-to-human affective conversations are also tedious and time-consuming. This paper describes construction of a multimodal affective conversational corpus based on TV dramas. The data contain parallel English-French languages in lexical, acoustic, and facial features. In addition, we annotated the part of the English data with speaker and emotion information. Our corpus can be utilized to develop and assess such tasks as speaker and emotion recognition, affective speech recognition and synthesis, linguistic, and paralinguistic speech-to-speech translation as well as a multimodal dialog system.

Keywords: Corpus construction, multimodal parallel data, affective conversation, television dramas

1. Introduction

Researchers have been working on spoken language processing for decades. Such technologies as speech recognition, speech synthesis, speech translation, and spoken dialog systems have been developed and progressed from a simple machine that responds to a small set of sounds to a more sophisticated artificial agent that can handle conversational speech. Unfortunately, most of these current technologies remain limited to recognizing what was said without addressing how it was said. For example, in conventional speech-to-speech translation, the verbal content of speech is translated, but its non-verbal content or paralinguistic information is ignored.

On the other hand, based on text, speech, and video, research on emotion recognition is gaining considerable traction in the fields of human-machine communication and multimedia retrieval (Schuller et al., 2009). Numerous official emotion recognition challenges (Schuller et al., 2009; Schuller et al., 2010; Schuller et al., 2011) have been held that improved the features and the classifiers that capture the traits of spoken emotions. Furthermore, different approaches, which involve the characteristics of sounds and prosody based on speaking styles and the expressions of emotional speech classification in anime films, have also been proposed (Hara and Itou, 2010). However, these studies only focused on non-verbal information for recognizing/classifying types of emotions without addressing verbal content.

Developing an artificial agent that mimics human interaction requires a speech-oriented interface that can handle both the verbal and non-verbal content often found in conversations. Unfortunately, much less work has examined the technologies that consider both matters, because performing such emotion-affected spoken language processing tasks is not trivial. Previous studies (Williams and Stevens, 1972; Picard, 1997; Murray and Arnott, 1993) reported that emotion largely changes acoustic realization, including pitch range, speech rate, voice quality, etc. Several approaches in emotional speech recognition (Mukaihara et al., 2017) and expressive speech synthesis (Tachibana et al., 2005) have attempted to enrich the models that include prosody and emotion information. However, these studies were mostly based on speech data that were read by professional actors. Affective communication is even more complex in cross-lingual situations because of expression differences in languages and cultures. Several studies (Anumanchipalli et al., 2012; Do et al., 2016; Kano et al., 2013) have recently attempted to translate paralinguistic information across different languages, but they remain based on speech read by bilingual speakers. As a result, natural conversation that includes the expression of emotions, which play an important role during human communication, has generally not been achieved yet by these systems.

Since the nature of data determines system's quality, the utilized data must have a gap that is as small as possible with real life emotion occurrences. The availability of multimodal, multilingual, and emotionally rich corpora is still limited. The tasks of recording and transcribing actual human-to-human affective conversations are also tedious and time-consuming. For such efforts, collecting affective conversational data acts as a starting point. This paper describes the construction of a multimodal affective conversational corpus based on recorded TV dramas that have already been broadcast. The data contain parallel English-French language with lexical, acoustic, and facial features. In addition, we annotated the part of the English data with speaker and emotion information.

2. Related Works

Several works are aiming to construct multimodal non-acted affective corpora. Douglas-Cowie et al. constructed

the HUMAINE Database, a multimodal corpus that consists of natural and induced data showing emotion in a range of contexts (Douglas-Cowie et al., 2007). Another is the SEMAINE Database, an emotion-rich conversational database, which was carefully constructed by recording interactions between Sensitive Artificial Listener (SAL) and users; each recording was transcribed and annotated with the actor's emotions (McKeown et al., 2012). Most databases were constructed with specific recording settings to obtain high-quality data. However, constructing such data is time-consuming and costly. Furthermore, these databases are mostly based only on monolingual transcription.

In multilingual corpora, the ATR basic travel expression corpus (BTEC) has served as the primary source for developing broad-coverage speech translation systems (Kikui et al., 2006). Its sentences were collected by bilingual travel experts from Japanese/English sentence pairs in travel domain phrasebooks. The ATR-BTEC has been translated into 18 languages, including French, German, Italian, Chinese, Korean, and Indonesian. Each language is comprised of 160,000 sentences. This corpus contains only text-based data. The Formosa Speech Database (Formosa) (Lyu et al., 2004), a multilingual corpus for Taiwanese-Hakka-Mandarin, was created by recording 49 hours of speech. Its corpus construction project took over one year to collect recordings from thousands of speakers. The constructed corpus consists of speech and text data. Recently, the Multi30K Database (Elliott et al., 2016), which is from Multilingual English-German Image Descriptions, was created for a WMT Shared Task of Multimodal Machine Translation. It is based on the Flickr30K Entities dataset (Plummer et al., 2015) that was selected and manually translated into German and French by human translators. However, this corpus also contains only image and text data.

On the other hand, several works have explored corpus construction from such existing data as video from movies or television. A conversation dialog corpus from movies and television (Nio et al., 2014) has been constructed to provide transcriptions of natural conversations of humans since the recording and transcription of actual human-to-human conversations are tedious and time-consuming to construct (Nio et al., 2014). Yasuhara et al. also constructed a large-scale multimodal dialog corpus from movies (Yasuhara et al., 2016). Their corpus, which consists of movie files and annotations that indicate the timing of dialog segments, contains 149,689 dialogue segments from 1,722 movies. Even though both corpora were constructed to provide patterns of human communication, they also only focused on monolingual data.

Compared to previous works, we construct a multimodal and multilingual affective conversational corpus from TV-series data. In addition, we annotated some of the English data with speaker and emotion information. Our corpus can be utilized for the development and the assessment of various tasks, such as speaker and emotion recognition, affective speech recognition and synthesis, linguistic and paralinguistic speech-to-speech translation as well as a multimodal dialog system.

3. Corpus Construction

3.1. TV-Series Data Resources

The corpus was constructed from American TV dramas which have already been broadcast and its DVD has been released in market. We used 40 episodes as resources. Each episode approximately consists of 600 utterances spoken by 40 to 60 speakers. The series were originally broadcast in English and have been dubbed into French. To construct parallel data, we utilized speech data which consist of their audio-visual data and text from both the original and French-dubbed versions of the TV series. The English-French data consist of parallel text subtitles, speech audio, and video images. A detailed overview of the resource can be seen in Table 1.

Total number of utterances	25,663
Average duration of each utterance	2 sec
Minimum duration of each utterance	0.8 sec
Maximum duration of each utterance	6 sec

Table 1: Raw resource data of TV series

3.2. Data Filtering

First, from the raw English text-resources, we selected the least noisy speech segments that were only spoken by a single speaker and removed the non-English speech utterances. Then we confirmed the accuracy of the subtitles and manually corrected them if they included typos. For easier processing in subsequent phases, we reformatted the timing information from an hour into a milliseconds format. The timing information from the newly reformatted transcription data were used to cut the speech audio into utterance-based segments.

Next, to construct parallel English-French data, we selected utterances in both languages that have identical timing and discarded the rest. Due to this process, the amount of resulting utterances was greatly reduced since the number of different-timed parallel utterances was quite high.

3.3. Feature Extraction

The feature extraction phase constructed a feature dataset of three modalities: acoustic, lexical, and facial cues.

- Lexical cues:
  We extracted the lexical features based on Google's Word2Vec (Mikolov et al., 2013). The Word2Vec model generates word-level features. After that, we calculated the average value of each attribute from each word of the utterances. Note that some utterances do not have corresponding lexical features with the names of places or people because they were not included in Word2Vec model's dictionary.

- Acoustic cues:
  We used openSMILE toolkit (Eyben et al., 2010) with feature configuration from the INTERSPEECH 2010 Paralinguistic Challenge. It consists of 38 low-level descriptors and 21 functionals that resulted in 1582 acoustic features.

- Facial cues:
  We used openFace toolkit (Baltrusaitis et al., 2016) to extract the facial features from the video. To generate utterance-based facial features, first we extracted the feature of each episode using openFace and then segmented the features by calculating the average value of each attribute of the features based on the utterance's timing information that was provided in the transcription data. Since actors in recorded productions/shows often move in various ways and positions from the screen, it was not possible to generate facial cues for such scenes if speaker faces were not captured on screen.

To avoid utterances with absence-features, we also synchronized the utterance-features to make sure that all the resulting utterances have all types of features. The synchronization proceeded by extracting each modality of the features, and then we automatically generated a list of utterances with every type of feature. The details of the process are shown in Fig. 1.

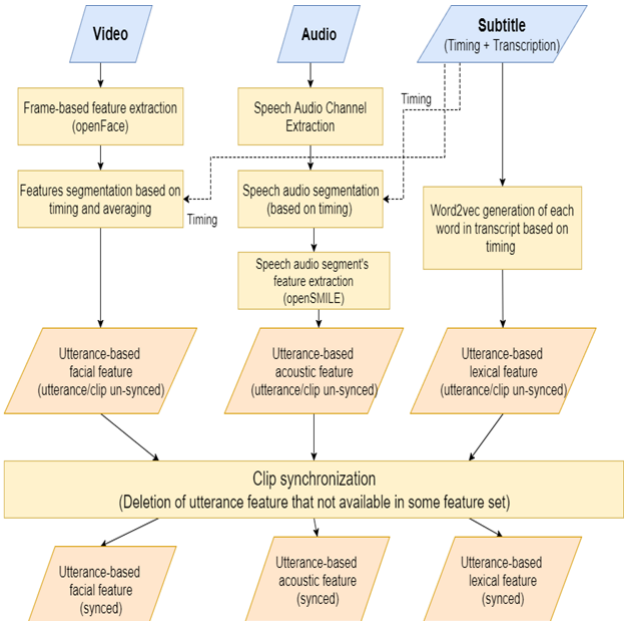

Figure 1: Multimodal feature extraction

4. Speaker and Emotion Annotation

In addition to the above multimodal dataset, we also annotated part of the English corpus with speaker and emotion information.

4.1. Speaker Annotation

Each utterance in the dataset is labeled with speaker information. The annotations were conducted manually by listening to and verifying the speaker of the utterances. The speaker label consists of 57 names of major characters who appeared in the TV series. We constructed a list of speakers by selecting those who made more than ten utterances in randomly selected episodes and appeared in more than one episode.

4.2. Emotion Annotation

We defined the emotion scope based on the circumplex model of affect (Cowie et al., 2011). Here, each utterance is labeled with emotion information based on its valence and arousal states. Valence measures the polarity of emotion; for example, 'happy' indicates a positive valence and 'sad' indicates a negative valence. Arousal measures the activity of an emotion; for example, 'tense' indicates high arousal and 'calm' indicates low arousal. To simplify the labeling, both valence and arousal were discretized into three labels: positive, neutral, and negative. The annotation for emotion was done manually by one person with a general trace program (GTrace) emotion annotation toolkit (Cowie et al., 2011) that consists of an emotion bar and a video screen. By using this toolkit, annotation was done by moving the bar's pointer to a location that corresponds to a particular emotion. Since GTrace resulted in real-valued annotation, we defined the ranges of the values or the thresholds for each class for each emotion measure. Then the annotation results by GTrace were classified based on the thresholds of the classes. Each utterance was labeled depending on the annotator's evaluation regarding the utterance's emotion. For example, an utterance that was made in an upset tone might be labeled as a 'negative' valence state and a 'positive' arousal state. Figs. 2 and 3 respectively show examples of valence and arousal annotation using Gtrace toolkit.

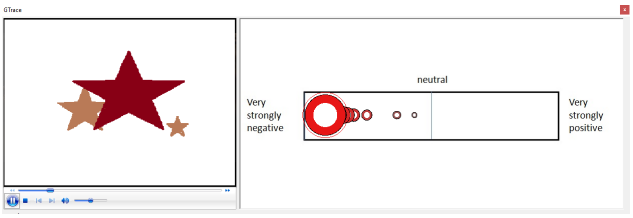

Figure 2: Valence annotation with Gtrace toolkit

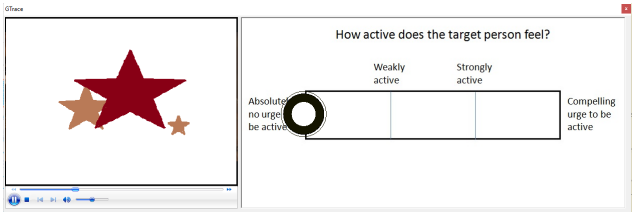

Figure 3: Arousal annotation with Gtrace toolkit

5. Corpus Analysis

The resulting corpus consists of 25,420 utterances that were spoken in English, 6,157 of which were annotated with speaker and emotion states. From the annotated utterances, 2,761 utterances have representations in acoustic, lexical, and facial cues. Among all the English-spoken utterances, only 6,114 have exact timing parallel utterances that were spoken in French.

Since affective communication in different languages and cultures might have differences in expressions, we analyzed

the constructed corpus to find them. Table 2 shows examples of English and French parallel utterances. Notice that parallel utterances don't have identical linguistic meaning. For example, the French utterance, *'c'est bien'*, should be *'it's good'* in English. In the TV series, the phrase is used by the speaker when certain work, which was done by other characters, is finished. Both utterances are used in the TV series in their respective language releases. The difference occurs because utterances in a TV drama form a conversation that is affected by the story's settings, characters, and language style. Even though the linguistic meaning is not identical, the parallel utterances have the same purpose or intention to be conveyed, and a cultural difference might affect the choice of words in the translation.

Timing (ms)	English	French
132348-133579	that's it	c'est bien
210927-212258	aye, but we have the wind	oui, mais on a le vent en poupe
215532-216624	you're coming home	tu rentres la maison

Table 2: Example of English-French parallel utterances

Next we analyzed the distribution of speakers and emotion types and focused on English data. Fig. 4 and Table 3 describe the distribution of the speaker class. In the constructed corpus, most speakers only made 10-25 utterances, while a particular speaker, probably that show's star, made over 200 utterances. As for emotion, Figs. 5 and 6 respectively illustrate the valence and arousal class distributions. Most of these utterances contain neutral emotions.

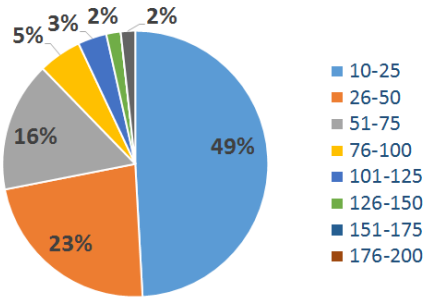

Figure 4: Speaker class distribution

Samples	Speakers
10-25	28
26-50	13
51-75	9
76-100	3
101-125	2
126-150	1
176-200	1

Table 3: Number of samples and speakers in data

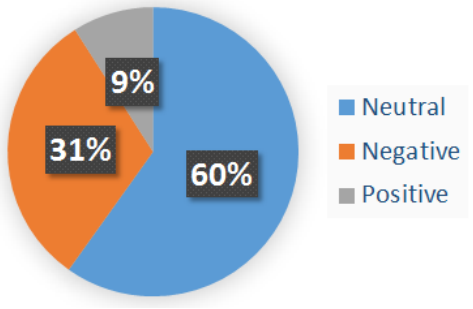

Figure 5: Valence class distribution

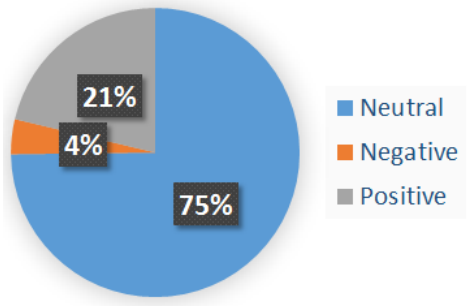

Figure 6: Arousal class distribution

6. Conclusion

In this work, we constructed a multimodal and multilingual conversational corpus from TV dramas. The data, which contain parallel English-French language in lexical, acoustic, and facial features, were annotated with speaker and emotion information. From our constructed corpus, even though we found that parallel speech may not have linguistically identical meaning, it still denotes the same thing or the same purpose. We can learn this by watching the video, although we may not realize it if we rely on speech itself for our understanding. In other words, such external speech factors as situation, cultural background, and speaker may affect word choices and speech meanings. We conclude that our corpus can be utilized to develop a paralanguage processing system that considers such factors. Future work will deepen our analysis of English-French speech data from the resources of TV-series data to increase the size of multilingual corpora since the current parallel utterances are only based on the exact timing of utterances in both languages.

7. Acknowledgement

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

8. Bibliographical References

Anumanchipalli, G., Oliveira, L., and Black, A. (2012). Intent transfer in speech-to-speech machine translation. In *Proc. of SLT*, pages 153–158.

Baltrusaitis, T., Robinson, P., and Morency, L. P. (2016). Openface: An open source facial behavior analysis

toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, March.

Cowie, R., Cox, C., Martin, J., Batliner, A., Heylen, D., and Karpouzis, K., (2011). *Issues in Data Labelling*, chapter Emotion-Oriented Systems: The Humaine Handbook, pages 215–244. Springer-Verlag Berlin Heidelberg.

Do, Q., Toda, T., Neubig, G., Sakti, S., and Nakamura, S. (2016). Preserving word-level emphasis in speech-to-speech translation. *IEEE Transactions on Audio, Speech and Language Processing*, 25:544–556.

Eyben, F., Wollmer, M., and Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1459–1462, New York, NY, USA. ACM.

Hara, Y. and Itou, K. (2010). Classification of emotional speech in anime films by using automatic temporal segmentation. In *Proc. of the second International Conference on Creative Content Technologies (CONTENT)*, pages 61–68, Lisbon, Portugal.

Kano, T., Takamichi, S., Sakti, S., Neubig, G., Toda, T., and Nakamura, S. (2013). Generalizing continuous-space translation of paralinguistic information. In *Proc. of INTERSPEECH*, pages 2614–2618.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Mukaihara, K., Sakti, S., and Nakamura, S., (2017). *Recognizing Emotionally Coloured Dialogue Speech using Speaker-Adapted DNN-CNN Bottleneck Features*, chapter Speech and Computer Lecture Notes in Computer Science, pages 632–641. Lecture Notes in Computer Science. Springer.

Murray, I. and Arnott, L. (1993). Toward the simulation of emotion in synthetic speech: a review of the listerature on human vocal emotion. *Journal Acoustical Society of America*, 93(2):1097–1108.

Nio, L., Sakti, S., Neubig, G., Toda, T., and Nakamura, S. (2014). Conversation dialog corpora from television and movie scripts. In *2014 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, pages 1–4, Sept.

Picard, R. (1997). *Affective Computing*. MIT Press.

Schuller, B., Steidl, S., and Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proc. INTERSPEECH*, pages 312–315, Brighton, United Kingdom.

Schuller, B., Steidl, S., Burkhardt, F., Devillers, L., Muller, C., and Narayanan, S. (2010). The INTERSPEECH 2010 paralinguistic challenge. In *Proc. INTERSPEECH*, pages 2794–2797, Makuhari, Japan.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. (2011). AVEC 2011 - the first international audio/visual emotion challenge. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 415–424, Memphis, Tennessee.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2005). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE*, 88:2484–2491.

Williams, C. and Stevens, K. (1972). Emotion and speech: Some acoustical correlates. *J. Acoust. Soc. Amer*, 52:1238–1250.

9. Language Resource References

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J., Devillers, L., Abrilian, S., Batliner, A., Amir, N., and Karpouzis, K. (2007). The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction*, ACII '07, pages 488–500, Berlin, Heidelberg. Springer-Verlag.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Kikui, G., Yamamoto, S., Takezawa, T., and Sumita, E. (2006). Comparative study on corpora for speech translation. 14(5):1674–1682.

Lyu, R., Liang, M., and Chiang, Y. (2004). Toward constructing a multilingual speech corpus for taiwanese (min-nan), hakka, and mandarin chinese. volume 9.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.*, 3(1):5–17, January.

Nio, L., Sakti, S., Neubig, G., Toda, T., and Nakamura, S. (2014). Conversation dialog corpora from television and movie scripts. In *2014 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, pages 1–4, Sept.

Plummer, B., Wang, L., Cervantes, C., Caicedo, J., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. volume abs/1505.04870.

Yasuhara, R., Inoue, M., Suga, I., and Kosaka, T. (2016). Large-scale multimodal movie dialogue corpus. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 414–415, New York, NY, USA. ACM.