An Intelligent Infrastructure Toward Large Scale Naturalistic Affective Speech Corpora Collection

Shreya G. Upadhyay[1]*, Woan-Shiuan Chien[1]*, Bo-Hao Su[1], Lucas Goncalves[2], Ya-Tse Wu[1], Ali N. Salman[2], Carlos Busso[2], Chi-Chun Lee[1]
[1]*Department of Electrical Engineering, National Tsing Hua University, Taiwan*
[2]*Department of Electrical and Computer Engineering, University of Texas at Dallas, USA*

***Abstract*—The field of *speech emotion recognition* (SER) aims to create scientifically rigorous systems that can reliably characterize emotional behaviors expressed in speech. A key aspect for building SER systems is to obtain emotional data that is both reliable and reproducible for practitioners. However, academic researchers encounter difficulties in accessing or collecting naturalistic large-scale, reliable emotional recordings. Also, the best practices for data collection are not necessarily described or shared when presenting emotional corpora. To address this issue, the paper proposes the creation of an *affective naturalistic database consortium* (AndC) that can encourage multidisciplinary cooperation among researchers and practitioners in the field of *affective computing*. This paper's contribution is twofold. First, it proposes the design of the AndC with a customizable-standard framework for intelligently-controlled emotional data collection. The focus is on leveraging naturalistic spontaneous recordings available on audio-sharing websites. Second, it presents as a case study the development of a naturalistic large-scale *Taiwanese Mandarin* podcast corpus using the customizable-standard intelligently-controlled framework. The AndC will enable research groups to effectively collect data using the provided pipeline and to contribute with alternative algorithms or data collection protocols.**

***Index Terms*—Speech Emotion Recognition, Database Consortium, Affective Computing**

I. Introduction

Affective computing is a rapidly growing field that aims to develop algorithms and systems in multidisciplinary research fields such as psychology, physiology, engineering, mathematics, and linguistics [1]. Given the ubiquitousness of speech-based solutions, the use of *speech emotion recognition* (SER) has gained attention in different fields, including customer satisfaction, healthcare systems, voice analytics, spoken dialogue systems, and social media analysis [2]–[4]. SER systems should be capable of characterizing emotional behaviors regardless of the context, target application, and recording conditions. The success of such systems heavily relies on the availability of high-quality corpora that are diverse, large-scale, and accurately labeled.

Data-driven machine learning approaches form the foundation of modern speech emotion recognition technology. Emotions, being complex psychological states that can vary across cultures and languages [5], require careful data collection to produce naturalistic datasets. These diverse datasets in different languages and domains are essential for developing accurate and reliable emotion recognition models that can generalize across linguistic boundaries [6]–[8]. Over the years, researchers have explored various methods of data collection to obtain more realistic datasets [9]–[13]. However, most of the available datasets have been collected in a "static" manner until the recent development of the MSP-Podcast dataset [14]. This dataset represents a more dynamic and naturalistic collection of emotional speech data, addressing some of the limitations of static datasets.

Constructing a large-scale naturalistic and reliable emotion dataset is a challenging task, particularly when it comes to maintaining a continuous flow of data that accounts for changing emotional expressions and contextual variations. The MSP-Podcast database [14] not only collects naturalistic speech affective data but also introduces a unique process that facilitates a continuous flow of data with controllable modules at each stage. For instance, consider the scenario of obtaining raw audio data from the wild, which undergoes multiple stages such as preprocessing, emotion retrieval, annotations, etc., and each of these steps is automated by integrating specific speech-based AI modules. Drawing inspiration from the earlier presentation of the MSP-Podcast database, this study aims to extract and present the process as a standardized pipeline and resources that can be openly accessed and utilized by the research community.

In data collection, not only the continuous in-flow of data is important but also the "controllability" over the quality and biases of data [14], [15]. This controllability not only enables transparency and reproducibility in the data collection process but also allows for customization by switching or tweaking various modules within the pipeline. This customization feature allows for the generation of context-specific datasets, addressing the challenge of applying SER across highly diverse contexts. Furthermore, the shown pipeline not only allows for controllable data collection but also incorporates intelligent module integrations to facilitate customized data collection. This is particularly important in dealing with the increasing scale of data (driven by the data-hungry nature of deep learning) and the complexity of audio data (due to its spontaneous nature), also, the continuous integration of evolving speech intelligence frameworks, such as improved speaker identification or fair retrieval techniques, can be seamlessly incorporated into the pipeline to enhance the quality and

This work was supported by the NSTC under Grants 111-2634-F-002-023, 110-2221-E-007-067-MY3, and the NSF under Grant CNS-2016719.
* All these authors contributed equally.

relevance of the collected data.

Considering these factors, this work presents our efforts in developing an infrastructure pipeline and demonstrates the usage of this process for collecting and releasing a version of affective speech media data in *Taiwanese Mandarin* language. Additionally, we have compiled a unified *affective naturalistic database consortium* (AndC) website that provides researchers and practitioners with access not only to the corpus, but also to the source code for the pipeline, models, example usages, and currently available datasets. Our aim is to spark interest and facilitate ease of usage for data utilization, data collection, and model development in this collective research endeavour, and to move towards a consortium. The paper is divided into two sections. The first section provides details on the contents of the AndC and offers a comprehensive guide to building affective datasets using the pipeline. The second section describes the collection of a large-scale naturalistic BIIC-Podcast corpus, which serves as a case study.

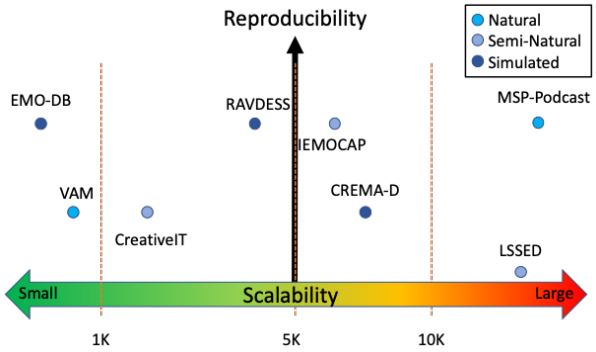


Fig. 1: The conceptual spectrum of widely used and latest SER corpora in the perspective of reproducibility and scalability. Scalability (x-axis) represents the amount of data, and Reproducibility (y-axis) depends on multiple factors, such as data collection settings, preprocessing details, and annotation process and so on.

II. Related Works

Since this paper contributes to developing a data collection infrastructure, our survey focuses on two main areas as insufficient openness in the process of data collection and unavailability of a structured pipeline framework for reliable dataset development.

A. Insufficient Transparency in Data Collection

For many years, scholars in the field of SER have been actively publishing speech emotion datasets. However, these datasets are often limited by their scale and contextualized settings. Recently, some large-scale naturalistic SER corpora have been released that helped to alleviate this bottleneck. For example, Fan et al. have published the LSSED [16], an English SER dataset containing 820 subjects and 206 hours of data that simulates real-world distribution. Further, the MSP lab has been continuously expanding its in the wild *podcast* dataset [14], which includes more than 237 hours of naturalistic speech recordings, and has also released a *conversational* type with 59 hours of data [17]. Despite these efforts, there are still challenges in creating emotional speech datasets that accurately represent the complexity and variation of emotions across different cultures and languages and also in preserving their naturalism.

Building and releasing datasets for SER requires considerations beyond just the data collection steps. In particular, scalability, reproducibility, and reliability are important factors to be addressed. Firstly, scalability should not only refers to the dataset size but also to its continuous alignment with real-time emotional variations in expressivity and perceivability. Secondly, reproducibility is crucial to evaluate the quality and re-implement procedures of the corpus and having transparency in producing the dataset can avoid related biases and quality issues. And, reliability demonstrates the robustness of a dataset, which is intertwined with scalability and reproducibility. Fig. 1 illustrates the position of some of the present emotional datasets on the scalability-reproducibility spectrum based on their naturalistic contents.

The lack of transparency and control in the data collection process of SER datasets poses a challenge to collaborating with practitioners in similar fields or domains. For example, while the LSSED [16] corpus by Fan et al. provides a large amount of data, it lacks transparency in the collection environment settings, making it difficult to replicate or collect similar recordings in other languages. Similarly, the MSP-Podcast [14], provides procedures but reproducing the results without the reference code is challenging. Therefore, a well-organized pipeline framework is needed, similar to Kaldi [18] in the speech domain, which provides a standard recipe for model evaluation and transparent pre-processing scripts. Such a framework facilitates research, enables easy reproduction of results, and promotes community contribution. Inspired by the success of Kaldi, we propose an AndC which contains an "easy-to-use" customizable intelligently-controlled framework for a large-scale naturalistic emotional speech database.

B. Unavailability of a Structured Pipeline Framework

There have been several proposals for data collection pipelines aimed at addressing the challenges of obtaining large-scale datasets for machine learning tasks. For example, in *natural language processing* (NLP), Leung's semi-automated pipeline [19] automatically sorts, reviews and discovers English articles, while Ning's CROWDAQ platform allows customizable and reusable data collection. In the EEG domain, Bigdely-Shamlo et al. [20] standardized early-stage processing, and Wang's automatic pipeline [20] evaluates multiple factors for intelligent robots using the visual SLAM algorithm. In the speech, Nagrani's fully automated pipeline [21] uses computer vision techniques to collect data from open-source media. Similarly in computer vision, Ding's automatic image-processing pipeline [22] distributes images to

multiple annotators simultaneously to optimize efficiency, and some annotation pipelines [23], [24] have also been presented.

Similar systematic frameworks have not yet been deployed in the SER fields, and none of the aforementioned pipelines avails any platforms or community that enables practitioners to easily reproduce them. To accelerate the development of SER, we propose the construction of a customizable intelligently-controlled framework that includes multiple efficiency analyses as preprocessing modules, emotion retrieval enablers, annotation distribution control, and model performance evaluation. To systematically and transparently aggregate all sources, we need a consortium that manages the details and also be sharable. Therefore, we organize an AndC website for calling collaborations in this field. This consortium aims to promote global collaboration within a customizable-standard framework that can be adapted to specific needs or requirements. Also, public monitoring of quality through the recipe can increase reliability, and standard procedures can also improve efficiency and transparency in the process. By releasing all components of the process involved in data collection, reimplementation hurdles can be reduced, and dataset searching can be made more convenient through a single consortium.

In the later sections, we elaborate on the complete data collection pipeline and present a case study corpus with a thorough analysis and evaluation of performances.

III. Affective naturalistic Database Consortium

This paper presents an *affective naturalistic database consortium* (AndC) website [1] which plays a pivotal role in providing a centralized platform for researchers to access the customizable intelligently-controlled pipeline framework and the naturalistic affective speech corpora that have been collected through this framework. To assist researchers with a convenient and user-friendly platform to access the framework and corpora effectively, the website offers a range of features, including detailed information about the speech recordings collection process, the selection of consented data, preprocessing procedures, and quality control measures that were implemented during the collection. This transparency in the process ensures the reliability and reproducibility of the corpora, also, can be used to produce corpora with different languages by customizing some components of the framework to efficiently adapt according to specific needs.

We divided the consortium website into two parts, *customizable-standard pipeline* and *affective speech corpora*. The website provides a comprehensive overview of the intelligent components of the standardized pipeline framework, allowing researchers to easily collect their own dataset based on specific criteria, such as speaker demographics, emotional categories, and audio quality. To increase the transparency in the data collection process, a preview function is added that enables researchers to listen to short audio clips demonstrating the effects of a few intelligent components, such as music filtering and SNR filtering. Additionally, researchers can access technical specifications like audio file format, sample rate, and bit depth to aid the upgrade in the pipeline. The website includes various resources for practitioners, such as code repositories, pretrained models, and forums for community discussion and collaboration. We offer a search function to help researchers find specific datasets based on keywords, affective states, or other criteria.

In order to make the pipeline more than just available, but also customizable, practitioners can add their own components to adapt to specific needs or requirements. For instance, if the data collection is for emotions in a different language, adjustments can be made to the *emotion retrieval* (ER) components to incorporate important linguistic features. To assist in this collaboration, we provide our pipeline and code to researchers via GitHub. The GitHub repository include all resources for researchers to replicate our data collection methodology. Additionally, we have structured our code modularly with documentation for each module. Each module is self-contained and allows switching between multiple components using a universal schema. We welcome the researchers or practitioners in the *affective computing* community to use our provided components or build and share their own code implementations, research findings and model components which they think are more effective for this kind of intelligent-governed data collection infrastructure, promoting a culture of collaboration and transparency in the field.

IV. Intelligently-Controlled Pipeline

The customizable intelligently controlled framework shown in Fig. 2 is divided into three broad stages to serve different purposes: preprocessing, emotion retrieval, and perceptual evaluations.

A. Enablers-based Pre-processing

The preprocessing stage of the data collection pipeline consists of two phases, namely, the audio preparation phase and the filter phase, which involve different intelligent enabler components. Once the raw audio collection is obtained, all the recordings are supposed to convert to a 16kHz, 16-bit, and 1-channel format using the Librosa toolbox [25] and then are passed through a *voice activity detector* (VAD) [26] to extract speech segments. Also, to eliminate utterances with multiple speakers, CountNet [27] is included to count the number of speakers in a "cocktail-party" scenario. After the audio preparation phase, the recordings are subjected to various filter phases to ensure high speech quality. These include a music filter and an SNR filter, etc. For instance, to filter music and noise in an utterance, an intelligent component such as music detector [28]and noise estimator [29] is utilized. Utterances with a music probability greater than and a *signal-to-noise ratio* (SNR) lower than a certain threshold are dropped. Only the utterances that meet all the criteria are passed to the next stage. As the pipeline is standard yet customizable, each phase's components can be modified, switched, or tweaked according to the specific requirements of the task at hand.

[1] http://www.andc.ai

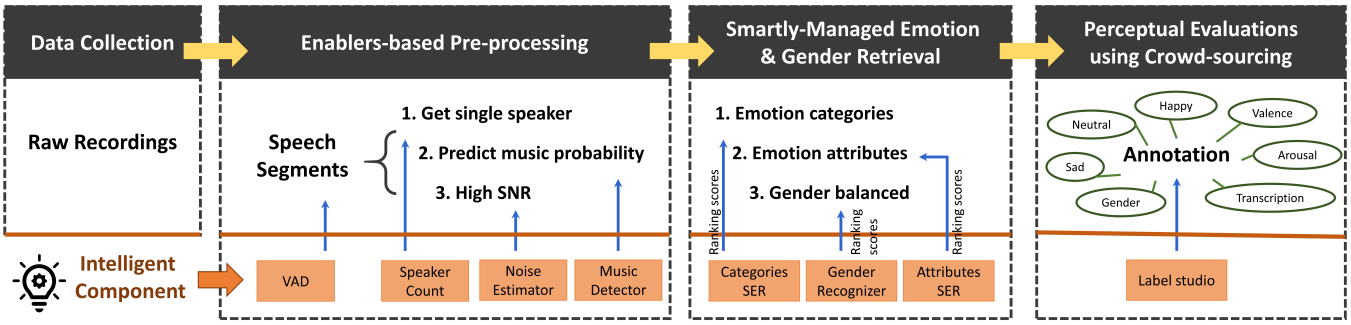


Fig. 2: Customizable Intelligently Controlled Pipeline Infrastructure.

B. *Smartly-Managed Emotion and Gender Retrieval*

After the pre-processing stage, the emotional retrieval step plays a crucial role in the data annotation process. Here, to create a well-distributed and unbiased emotional dataset, emotion detectors are used to retrieve emotional states from an unlabelled pool of speech recordings and rank utterances based on their emotional content. This process is similar to the approach used to retrieve emotional behaviours in the MSP-Podcast database [15]. Apart from the dominant emotions, one of the main difficulties in creating a comprehensive emotional corpus is acquiring emotions that are typically more difficult to elicit, such as *Fear* and *Disgust*. These emotions are often more difficult to capture in non-controlled and non-acted environments, making it essential to employ a mix of emotion prediction models that were trained with corpora that contain the harder-to-obtain emotional states ([15], [30], [31]). By targeting these minor emotions with a low and high range of attribute predictions, the data collection process can be continuous but still controlled. These emotion prediction models act as intelligent components that facilitate the emotional retrieval process.

For the emotional retrieval step, each intelligent component processes each utterance received and assigns a probability score for emotion classification or a numerical value for emotional attributes associated with the input utterances. The scores/values generated by each component are then recorded into a score sheet for each model, which aggregates all the processed utterances. Based on the desired emotional state we want to retrieve for annotation, the score sheet is sorted from highest to lowest or lowest to highest. Also, having a record of emotional scores/values predicted by each component, we use it to validate our model performance compared to actual ratings received from annotators' perceptual evaluations. This helps to identify bad-performing models and make adjustments to improve the performance or intelligence of the models.

This pipeline framework not only targets emotional content but also determines the gender of each unlabelled speech utterance. An intelligent component is utilized to predict gender and control gender balance. In summary, the retrieved emotional and gender predictions are ranked using scores from all components to prioritize high emotional content and minority emotional states. This ranking helps set thresholds to determine which gender and emotional states should be prioritized for annotation, ensuring a more comprehensive and accurate annotated dataset while minimizing bias.

Fig. 3: Assessment for the emotional labeling of the corpus.

C. *Perceptual Evaluations using Crowd-sourcing*

After the emotion retrieval step, the emotional utterances undergo a perceptual evaluation using *label studio* [32]. This stage involves human annotating the utterances with emotional attributes (*Arousal*, *Valence*, *Dominance*) and categorical emotions (*Happiness*, *Anger*, *Sadness*, etc.), which is a common approach used in many existing affective data collections [12], [33]. We followed a similar approach in our pipeline.

The questionnaire used for annotation is displayed in Fig. 3. The emotional utterances retrieved from the previous stage of the pipeline are annotated on a 7-point Likert scale to evaluate *Valence* (ranging from very negative to very positive), *Arousal* (ranging from very calm to very active), and *Dominance* (ranging from very weak to very strong) using (Q.2-4) of the

Fig. 3. To assist the evaluators in annotating these dimensional attributes, we employ *self-assessment manikins* (SAMs) [34], [35] as a visual guide. In the questionnaire (Q.5-6) of the Fig. 3, evaluators are asked to select one primary emotion that they perceive best characterizes the emotional utterance from a list of eight primary emotions: *Anger*, *Sadness*, *Happiness*, *Surprise*, *Fear*, *Disgust*, *Contempt*, and *Neutral*.

Naturalistic speech recordings involve real-world communication and it is challenging to elicit emotional states that cannot be adequately expressed with a single emotion. So, we also annotate for secondary emotions similar to Busso et al. [33], where the evaluators can select all the possible emotional states they perceive in utterances (e.g., *Anger* + *Depressed* + *Annoyed*). Here, the list of secondary emotional states includes *Amused*, *Frustrated*, *Depressed*, *Concerned*, *Disappointed*, *Excited*, *Confused*, and *Annoyed*. To reduce the cognitive load, similar emotional categories are grouped together. In addition to annotating emotional states, we also annotate utterances for correct transcription and speaker gender.

V. BIIC-Podcast: A case study

This section of the paper presents a case study on the BIIC-Podcast, which is an emotional database in *Taiwanese Mandarin* language. It contains two sections, one including a detailed description of the database collection and its evolution over time, and another for the analyses and SER experiments after the collection of the database.

A. Database Collection and Evolution

The BIIC-Podcast database is the collection of emotional speech recordings in *Taiwanese Mandarin* language. The audios are gathered from various audio-sharing platforms that provide *creative commons* (CC) licenses under CC-BY or CC-BY-SA. To ensure diversity in the collection, we carefully selected topics (sports, lifestyle, business, music, and more), including monologues and conversations, and in various categories such as drama, interviews, casual conversations etc. We also maintained a balance between male and female speakers and limited the duration of one speaker to prevent speaker bias.

The raw audio recordings included in the BIIC-Podcast database have undergone all the stages of the intelligently-controlled pipeline shown in Section IV. This corpus is continuously collected in an ongoing process through the mentioned pipeline, which involves switching or tweaking some of the components to include language-specific emotional information. The choice of the number of annotators is subjective and depends on the needs and requirements of the specific work. For this data collection, we opt to have a group of 3 to 5 annotators to rate the emotional labels for each single utterance. Fig. 4 shows the corpus size progression report of recordings collected for the BIIC-Podcast over the course of 11 months. The data collection of BIIC-Podcast, from April-22 to Feb-23, reaches 60k available naturalistic emotional utterances. According to Fig. 4, there is a clear and steady upward trend in data size, with a rapid exponential increase

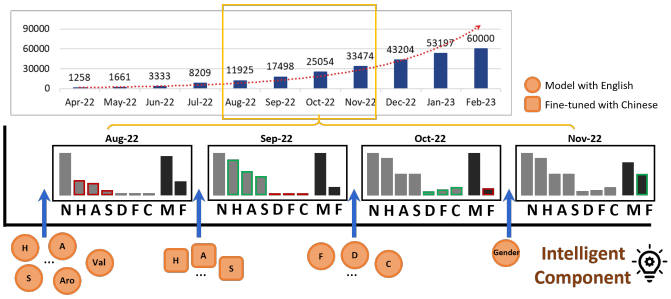


Fig. 4: Corpus evolution chart over a collection period.

in data quantity occurring around the fifth month, i.e. Sep-22. This increase in corpus size with balanced attributes is the result of a customized and well-controlled pipeline, which incorporates enabler-based prepossessing, intelligent emotion retrieval components, and perceptual evaluation techniques. This data collection evolution makes the infrastructure highly promising for the development of large-scale naturalistic affective corpora.

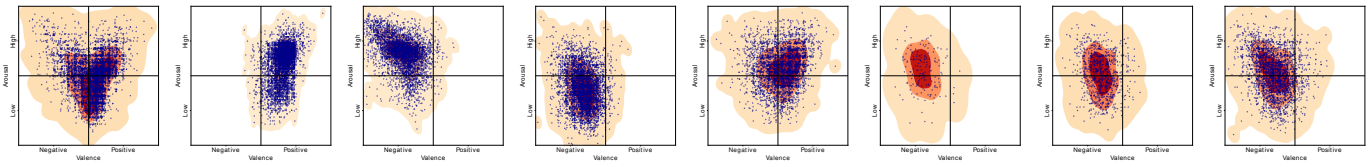
The lower section of Fig. 4 visually illustrates the significant progression of the BIIC-Podcast dataset, showcasing the occurrence of switching or tweaking of intelligent components during the data collection process. For instance, in Aug-22, there were fewer samples of *Happiness*, *Anger*, and *Sadness*. To address this, we added new fine-tuned retrieval models using *Taiwanese Mandarin* data specifically for these emotions, in order to capture language-specific features. As a result of this customization, an increase in the number of samples for these emotions is seen in the subsequent month.

B. Analysis and Experimental Study of the Database

1) Perceptual and Acoustic Analyses: This subsection of the paper involves analyses of the collected BIIC-Podcast database to examine the diversity and variability of emotions present in the database.

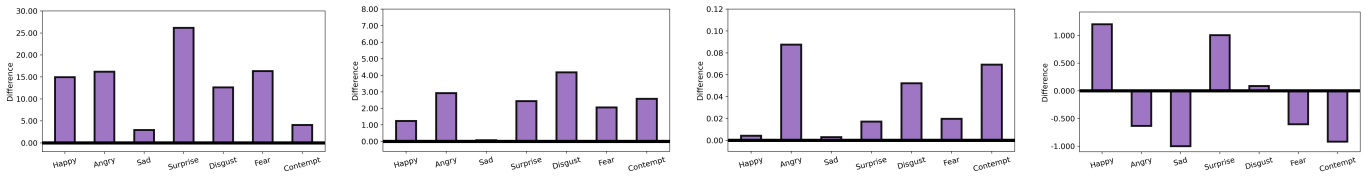
Emotional Diversity: Here, we assess the diversity of emotional content present in the BIIC-Podcast. The final emotional labels are the aggregates of the assigned labels by the workers using the majority vote rule. The distribution of emotions in the BIIC-Podcast is shown in Fig. 7. From Figs. 7b, 7c, and 7d, it is evident that there is a relatively balanced distribution of *Valence*, *Arousal* and *Dominance* in the dataset. The categorical emotion distribution, as shown in Fig. 7a, also exhibits diverse emotions with categories such as *Neutral*, *Happiness*, *Anger*, and *Sadness* having a larger number of samples. The database includes instances of *Fear* and *Disgust*, although these are relatively fewer compared to other more common and expressive emotions present. These emotions are challenging to capture in naturalistic interactions. Additionally, there also exists some samples that do not reach a consensus among the workers.

Reliability of Emotional Annotations: Table I shows the reliability of emotional annotations obtained with the perceptual evaluation of the raters recruited by our laboratory. For primary



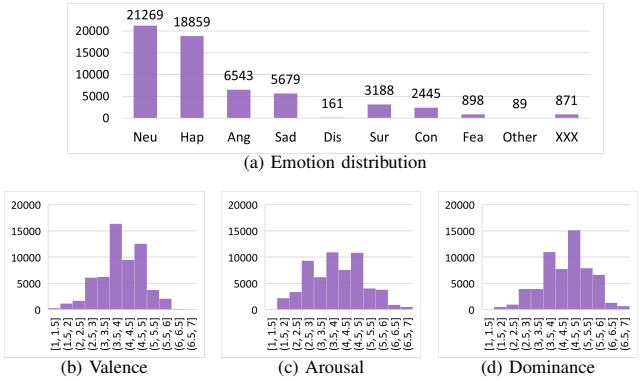
(a) Overall (b) Happiness (c) Anger (d) Sadness (e) Surprise (f) Disgust (g) Fear (h) Contempt

Fig. 5: Primary categorical emotions of BIIC-Podcast corpus on Valance-Arousal (VA) axis.



(a) F0 (Hz), speaker z-norm (b) Intensity (dB), session z-norm (c) Energy (dB), session z-norm (d) Speech Rate (# phones/sec)

Fig. 6: Relative differences toward *Neutral* emotion of acoustic features for BIIC-Podcast emotion categories.



(a) Emotion distribution

(b) Valence (c) Arousal (d) Dominance

Fig. 7: Distribution of emotion.

emotions, we measure *inter-annotator agreement* (IAA) using Fleiss' kappa (κ), reaching $\kappa = 0.337$. For attribute-based annotations, we assess Krippendorff's alpha coefficient (α) to evaluate the inter-annotator agreement, which yielded values all above 0.4.

Categorical Emotions on Valance-Arousal (VA) Axis: Fig. 5 displays the scatter plot of all primary categorical emotions of the BIIC-Podcast corpus on the VA plane. The plot is estimated by calculating the average distance to the 20 closest neighboring samples from a given point. This analysis aims to determine whether specific emotional samples are located in their expected quadrants or not, as described in the literature [36], [37]. As shown in Fig. 5, the BIIC-Podcast's emotional samples are mostly distributed in their expected quadrant, but also slightly scattered. This finding may be due to the complexity of emotions present in naturalistic recordings.

Relative Differences with Acoustic Features: Acoustic features like fundamental frequency (f0), intensity, energy, and speech rate are critical for analyzing and understanding speech and the emotions conveyed through it. We compare the relative acoustic differences between emotional and neutral sentences to assess if they follow the expected trends. Fig. 6 illustrates

TABLE I: Reliability of annotations.

	Caegorical Emotion (κ)	Valence (α)	Arousal (α)	Dominance (α)
IAA	0.337	0.461	0.418	0.432

the variations in acoustic features of BIIC-Podcast samples across emotion categories relative to the *Neutral* emotion. We extract fundamental frequency (F0), intensity, and energy, using the Praat tool [38]. For the speech rate shown in Fig. 6d, we estimated the average number of spoken phonemes per second for each emotion and the phone estimation is based on the forced alignment results. From Fig. 6a, we observe that *Sadness* and *Disgust* have low relative differences compared to other emotions and the relative differences in intensity shown in Fig. 6b conform to the trends presented in literature [37], [39], [40]. The energy and speech rate plots in Fig. 6c and 6d also align with the expected behaviour shown in the literature [37]. These insights suggest that this large-scale BIIC-Podcast conforms to the expectation of the emotion-specific modulations.

2) Data SER Modelling: This subsection presents the results of SER modeling experiments using the BIIC-Podcast corpus to examine its performance in the SER context.

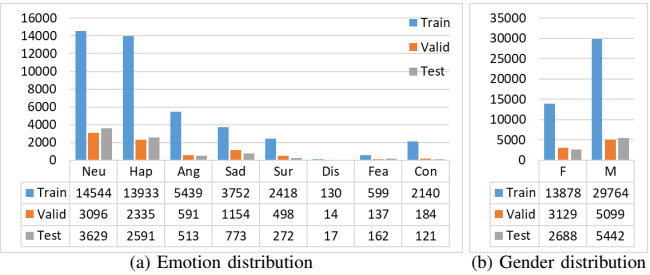
Experiment Settings: For all the experiments, the *Adam* optimizer is utilized with a learning rate and decaying factor of 0.0001. For the 4-category emotion (i.e., *Neutral*, *Anger*, *Happiness*, and *Sadness*) classification SER systems, the models are trained using back-propagation with the cross-entropy loss function, while the regression SER tasks use *mean squared error* (MSE) loss. The network is trained for a maximum of 50 epochs, with a batch size of 64 and early stopping is applied. The evaluation metrics used for the SER categorical classification and regression tasks are *unweighted average recall* (UAR) and *concordance correlation coefficient* (CCC) respectively. The models are trained and evaluated on the Tran/Valid/Test splits, which are detailed in section V-B2.

Train/Valid/Test Splits: In BIIC-Podcast corpus, we split the

TABLE II: SER Performance table with different functional, low level descriptors, and self-supervised features considering different sequential and non-sequential architecture models for *Categorical emotions*, *Arousal*, *Valence*, and *Dominance*.

	Categorical Emotions (UAR (%))			Arousal (CCC)			Valence (CCC)			Dominance (CCC)		
	eGeMAPS	ComParE	wav2vec	eGeMAPS	ComParE	wav2vec	eGeMAPS	ComParE	wav2vec	eGeMAPS	ComParE	wav2vec
DNN	37.34	34.67	41.10	0.361	0.358	0.368	0.334	0.343	0.352	0.362	0.353	0.371
CNN	38.69	36.01	43.36	0.399	0.386	0.372	0.335	0.347	0.357	0.369	0.361	0.385
LSTM	37.23	36.26	43.55	0.403	0.412	0.394	0.366	0.357	0.373	0.379	0.365	0.391
Transformer	-	-	44.86	-	-	0.426	-	-	0.381	-	-	0.397

data into predefined training, validation, and testing sets, maintaining the distribution of emotions and genders. This can help us avoid any biases towards a particular emotion or gender during the model training and testing phases Fig. 8a and 8b depict the distribution of the train, validation, and test sets based on emotions and gender. For speaker separation, we allocate distinct podcast series for each set to reduce the number of shared speakers across the splits.



(a) Emotion distribution (b) Gender distribution

Fig. 8: Distribution of BIIC-Podcast Train/Valid/Test sets.

SER Modelling: To have an insight into initial SER model performance with BIIC-Podcast, various combinations of functional, low-level descriptors (eGeMAPS, ComParE), and self-supervised pretrained embedding (wav2vec 2.0) features are used along with different sequential (CNN, LSTM, Transformer) and non-sequential (DNN) model architectures. The model is trained for *Categorical Emotions*, and for the emotional attributes of *Arousal*, *Valence*, and *Dominance*. This analysis aims to assess the performance of BIIC-Podcast on various combinations of features and architectures for the SER tasks. The performance results are recorded in Table II.

The results of the SER experiments, as shown in Table II, provide insights into the initial performance of various models trained on the BIIC-Podcast corpus. These results can serve as reference performance values for future research in the development of SER models using this corpus.

VI. Discussion and Conclusion

This paper serves as an effort towards the *affective naturalistic database consortium* (AndC) for collecting large-scale naturalistic affective speech databases. The goal of AndC is to provide access not only to emotional databases but also to a customizable, intelligently-controlled pipeline infrastructure for collecting naturalistic emotional speech recordings in a "continuous" and "controllable" manner. This intelligent infrastructure ensures reliable data collection by enabling high-level control over emotional distribution through its integrated intelligent components at each pipeline stage. Furthermore, the customizable setting of the pipeline opens up a broad possibilities for collecting data in diverse languages.

To showcase the practical application of the proposed intelligent infrastructure, we have collected and released the BIIC-Podcast, an afftective *Taiwanese Mandarin* dataset in conjunction with this paper. After the collection of BIIC-Podcast database, the database has been analyzed and experimented to understand its emotional characteristics, thus highlighting the effectiveness of the proposed pipeline in advancing research in *affective computing*. The results and analyses from sections V-B2 and V-B1 demonstrate that the pipeline is successful in controlling the emotional information during the continuous data collection process.

Overall, this paper provides an efficient and reliable method for collecting naturalistic affective speech corpora, which is crucial for advancing research in *affective computing*. The future work is the continue efforts to expand the AndC and increase its diversity in terms of languages, and emotional states. In addition, we will focus on the development of additional methods to assess the pipeline's efficacy in reducing bias [41] and improving the representativeness of emotions.

Ethical Impact Statement

The release of our *affective naturalistic database consortium* (AndC) aims to advance affective research forward and drive the development and availability of emotionally rich databases in multiple languages. The AndC's customizable intelligent pipeline incorporates important measures for controlling emotions in continuous affective data collection. Our approach offers opportunities for researchers or practitioners to contribute and promotes transparency in data collection practices. We acknowledge that there may be areas that our process may have overlooked, and with the public release of the AndC, we encourage other researchers to share their input and collaborate in further improvements in the pipeline.

Aside from making the AndC available to the public, we are also releasing the BIIC-Podcast, a *Taiwanese Mandarin* affective database. This corpus provides an opportunity for researchers to evaluate their methods in the context of *Taiwanese Mandarin* language and may help mitigate biases in research as many large-scale naturalistic affective speech databases are predominantly available in *English* language.

References

[1] Resham Arya, Jaiteg Singh, and Ashok Kumar, "A survey of multidisciplinary domains contributing to affective computing," *Computer Science Review*, vol. 40, pp. 100399, 2021.

[2] Laurence Devillers, Christophe Vaudable, and Clément Chastagnol, "Real-life emotion-related states detection in call centers: a cross-corpora study," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017, NIPS'17, p. 6000–6010, Curran Associates Inc.

[4] J.C. Acosta, *Using emotion to gain rapport in a spoken dialog system*, Ph.D. thesis, University of Texas at El Paso, El Paso, TX, USA, December 2009.

[5] Nangyeon Lim, "Cultural differences in emotion: differences in emotional arousal level between the east and the west," *Integrative medicine research*, vol. 5, no. 2, pp. 105–109, 2016.

[6] Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn, "Cross-lingual word embeddings for low-resource language modeling," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 937–947.

[7] Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko, "How translation alters sentiment," *Journal of Artificial Intelligence Research*, vol. 55, pp. 95–130, 2016.

[8] Kaisla Kajava, Emily Öhman, Piao Hui, and Jörg Tiedemann, "Emotion preservation in translation: Evaluating datasets for annotation projection," *Proceedings of Digital Humanities in Nordic Countries (DHN 2020)*, 2020.

[9] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al., "A database of german emotional speech.," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.

[10] Steven R Livingstone and Frank A Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, pp. e0196391, 2018.

[11] Milan Gnjatović and Dietmar Rösner, "Inducing genuine emotions in simulated speech-based human-machine interaction: The nimitek corpus," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 132–144, 2010.

[12] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[13] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan, "The vera am mittag german audio-visual emotional speech database," in *2008 IEEE international conference on multimedia and expo*. IEEE, 2008, pp. 865–868.

[14] Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[15] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[16] Weiquan Fan, Xiangmin Xu, Xiaofen Xing, Weidong Chen, and Dongyan Huang, "Lssed: a large-scale dataset and benchmark for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 641–645.

[17] L. Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso, "The MSP-conversation corpus," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 1823–1827.

[18] D. Povey, A. Ghosha, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and J. Silovsky, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, Waikoloa, HI, USA, December 2011.

[19] T. Leung and L.N. Perkins, "Counting protests in news articles: A dataset and semi-automated data collection pipeline," *ArXiv e-prints (arXiv:2102.00917)*, pp. 1–7, February 2021.

[20] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A Robbins, "The prep pipeline: standardized preprocessing for large-scale eeg analysis," *Frontiers in neuroinformatics*, vol. 9, pp. 16, 2015.

[21] A. Nagrani, J.S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 2610–2620.

[22] H Jane Ding, Catherine M Oikonomou, and Grant J Jensen, "The caltech tomography database and automatic processing pipeline," *Journal of structural biology*, vol. 192, no. 2, pp. 279–286, 2015.

[23] Agrim Gupta, Piotr Dollar, and Ross Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5356–5364.

[24] Damen Dima et al., "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *International Journal of Computer Vision*, vol. 130, no. 1, pp. 33–55, 2022.

[25] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8, pp. 18–25.

[26] John Wiseman and Ivan Yu Bondarenko, "Python interface to the webrtc voice activity detector," 2016.

[27] Fabian-Robert Stöter, Soumitro Chakrabarty, Bernd Edler, and Emanuël AP Habets, "Countnet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 268–282, 2018.

[28] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," *arXiv preprint arXiv:1703.01789*, 2017.

[29] Aaron Nicolson and Kuldip K Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, 2019.

[30] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, October-December 2014.

[31] S.R. Livingstone and F.A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, May 2018.

[32] Andrey Holmanyuk Maxim Tkachenko, Mikhail Malyuk and Nikolai Liubimov, "Label Studio: Data labeling software," 2020-2022, Open source software available from https://github.com/heartexlabs/label-studio.

[33] Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso, "The msp-conversation corpus," *Proc. Interspeech 2020*, pp. 1823–1827, 2020.

[34] Lorenz Fischer, Dieter Brauns, and Frank Belschak, "Zur messung von emotionen in der angewandten forschung," *Beiträge zur Wirtschaftspsychologie*, 2002.

[35] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[36] James A Russell and Geraldine Pratt, "A description of the affective quality attributed to environments.," *Journal of personality and social psychology*, vol. 38, no. 2, pp. 311, 1980.

[37] W.-S. Chien, S. Upadhyay, W.-C. Lin, Y.-T. Wu, B.-H. Su, C. Busso, and C.-C. Lee, "Monologue versus conversation: Differences in emotion perception and acoustic expressivity," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2022)*, Nara, Japan, October 2022, pp. 1–7.

[38] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[39] Klaus R Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[40] Roddy Cowie and Randolph R Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[41] L. Martinez-Lucas, A. Salman, S.-G. Leem, S.G. Upadhyay, C.-C. Lee, and C. Busso, "Analyzing the effect of affective priming on emotional annotations," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, September 2023.