Estimation of Hazardous Environments Through Speech and Ambient Noise Analysis

Andrea Veronica Porco
Dept. Information Engineering
University of the Ryukyus
Nishihara, Japan

Kang Dongshik
Dept. Information Engineering
University of the Ryukyus
Nishihara, Japan

***Abstract*—In recent years, significant attention has been directed towards the development of artificial empathy within the engineering academic community. Replicating artificial empathy necessitates the capability of agents to discern human emotions and comprehend environmental risks. Analyzing acoustic data in real environments offers a higher level of non-invasive privacy compared to video and camera data, limiting the agent's understanding to specific patterns. However, current studies are negatively affected by subjective inferences from real data, which can result in inaccurate predictions, leading to both false positives and negatives, especially when contextual data and human speech are involved. This paper work proposes the estimation of a dangerous environment in accordance with the emotional speech and additional ambient noises. In this approach we implement a variational autoencoder model in conjunction with a classifier for training the classification task. Additional regularization techniques are applied to bridge the gap between the original training data and the expected data. The classifier utilizes feature data generated by the variational autoencoder to extract class patterns and determine whether the environment is hazardous. Emotional speech is classified as angry, sad, or scared emotions, contributing to the classification of danger, while happy, calm, and neutral emotions are considered safe. Various ambient noise types, including gunfire and broken glass, are categorized as dangerous, while real-life indoor noises like cooking, eating, and movements are considered safe.**

***Keywords*—Dangerous environment detection; speech analysis; acoustic audio analysis; ambient noises; variational autoencoder model; empathetic systems**

I. Introduction

Ensuring the safety of individuals within indoor environments is a paramount concern, with implications spanning from residential spaces to critical infrastructure. The ability to accurately assess and respond to potential hazards is crucial for safeguarding lives and minimizing risks. In recent years, the pursuit of artificial empathy within the engineering domain has gained significant traction, aiming to imbue computational agents with the capacity to comprehend human emotions and navigate environmental dangers. An avenue of exploration in this pursuit involves the analysis of acoustic data, providing a non-intrusive means of understanding the surrounding environment. Unlike more invasive data sources like video and cameras, acoustic data analysis preserves privacy by focusing on discernible patterns, presenting a valuable approach for ensuring security in various settings.

The research presented in this paper addresses a critical facet of safety by proposing a method for estimating hazardous environments through the evaluation of emotional speech and ambient noises. This approach not only advances the field of artificial empathy but also holds substantial promise for real-world applications, particularly in indoor acoustic analysis and speech classification. The implications of this work extend beyond the academic realm, offering tangible benefits for society at large. The ability to accurately classify emotions and distinguish between safe and hazardous sounds has significant societal impact, enhancing security in public spaces, homes, and workplaces. Moreover, in the realm of engineering, the proposed method contributes to the refinement of hazard detection systems, with potential applications in areas such as smart home technologies, security surveillance, and other safety-critical environments. This research sets the stage for a more nuanced understanding of the acoustic environment, bridging the gap between subjective inferences and objective safety assessments, thereby paving the way for advancements in both theoretical understanding and practical implementation.

Analyzing a diverse range of events, including human speech and ambient sound, presents a formidable challenge for artificial agents. Consequently, accurately judging environmental characteristics becomes a complex endeavor. Moreover, this task necessitates numerous sensors, such as cameras for image and video processing, coupled or decoupled infrared sensors, and other costly apparatus, making the practical implementation of artificial home assistants exceedingly challenging to achieve. Addressing the challenges inherent in enhancing their practical implementation involves different tasks, such as managing real-time processes effectively, as evidenced in related papers [1]–[3].

Another challenge is to ensure accurate object localization such as in [4]–[7] where they proposed for example, a convolutional recurrent neural network for joint sound event localization and detection of multiple overlapping sound events in three-dimensional space. In particular the sound event localization and detection is extensively utilized by works based on robotics navigation and natural interaction with surroundings.

Background noise treatments and reduction have been extensible studied such as in [8]. The reduction of the noise comes to fulfil two different targets, the human hearing safety and the reduction of background noise to interpret another sounds or a clear speech. The source separation of overlapped sounds in acoustic event identification have studied in [9] to feat one of the pending challenges. While in [10]–[13], a study of event detection by ambient sounds analysis was performed, trying to give realism to the scenario through the addition of diverse types of ambient sounds.

Among several challenges, to perform a precise and realistic danger classification and estimation is especially required, considering subjective human perspectives and the critical task of minimizing the false alarms [14]. Some studies were carried trying to estimating hazardous environment from ambient sounds with support vector machine models such as in [15]. However the issue continue active and open to date. The significance of false alarms cannot be overstated, owing to the inherent subjectivity found in real-life scenarios. Environmental sounds has been under-researched compared to standard speech and music, and its understanding tend to be subjective depending on the scenario and the listener [16].

Previous research endeavors have explored various ambient sounds as cited in [17], yet remarkably, the emotional states of individuals within the room have never been integrated into the equation, amplifying the complexity of the challenge at hand, and making it more realistic.

Furthermore, the usage of generative models have increased in the study of human emotions and context analysis due to the flexibility and versatility to represent and analize different types of data present all together in the same audio frame [18]–[24]. Generative models, particularly when integrated with classifiers such as variational autoencoders (VAEs), prove to be highly advantageous for classification tasks in the domain of speech and also with ambient sound. The combination of generative and discriminative capabilities allows for effective feature extraction and representation learning, enhancing the model's ability to discern patterns in complex audio data. The limitations of generative models in this context are primarily associated with tasks that demand perfect data reconstruction. Challenges arise when attempting to faithfully reproduce the intricate details of diverse audio signals, including variations in speech patterns, accents, and environmental sounds. However, in classification tasks, where the focus is on discerning relevant features rather than achieving precise data reconstruction, these limitations are mitigated. The flexibility and adaptability of generative models make them well-suited for classification applications, offering a powerful and efficient approach to audio analysis.

To the best of our knowledge, the detection of a dangerous environment was never judged by an emotional speech analysis in combination with ambient noises analysis with generative models, such as a variational autoencoder (VAE) model that learn the characteristics related with a subjective environment. Additionally, the proposed model make an adjustment of the difference among input data and expected data with phonetic and prosody features.

II. Proposed Approach

A. Proposed Model

The proposed model falls under the category of semi-supervised learning, which is a hybrid method combining labeled and unlabelled data. In this approach, the classifier learns from labeled examples and also utilizes information from unlabelled data to enhance its performance. Within our model, the VAE serves a dual purpose: it functions as an unsupervised autoencoder, learning a condensed representation of the data, and as a supervised classifier, predicting emotional classes. This dual role is possible because the model incorporates both the reconstruction loss (unsupervised) and the danger classification loss (supervised), allowing it to harness the advantages of both labeled and unlabelled data. In essence, our model is semi-supervised because it integrates labeled emotional class data along with unlabelled Mel spectrogram data during the training process, optimizing its performance.

The proposed classifier model utilizes the latent space generated by the VAE model. In our approach, these two models operate independently; first, the VAE is trained separately, and then the pre-trained encoder from the VAE, which has learned from unsupervised data, functions as a feature extractor in the classifier. This encoder transforms the data into a compact representation, which is then processed through a classifier. This classifier is specifically trained to predict danger labels based on these encoded features, which are both from labeled and unlabelled data. Consequently, our model is categorized as a semi-supervised learning approach incorporating both types of data during its training process.

Moreover, the regularization task in the extended VAE provides additional control over the decoded data. This control is achieved by utilizing pre-processed data and its representations within the VAE model, allowing for a more refined and controlled learning process [25]–[28].

The phonetic and prosody characteristics of each value derived from the input and decoded data, will be compared. The aim is to ensure that the phonetic and prosody attributes of the VAE representation closely match the pre-processed values from the input data during the extended VAE training process.

In this approach, the classifier relies on the trained encoder for its classification tasks. Consequently, the regularization process also impacts the final results of the classifier. Classifying the extended dataset containing dangerous patterns presents a challenge for our classifier. This challenge arises not only from the variations in volume and intonation between the actors' utterances but also from the inclusion of surrounded noises specific to the content of danger.

The prosodic control and regularization was previously observed in [29], [30]. The phonetic and prosodic features considered include spectral bandwidth, spectral contrast, and formants from 1 to 5 extracted from each input audio's Mel spectrogram. Our proposed classifier, coupled with the phonetic and prosodic regularized VAE method, encompasses multiple tasks. The subsequent subsections will delve into the specifics of each task involved in our approach.

The basic architecture of the hazardous environment classifier model can be observed in Fig. 1. The variational autoencoder is represented with an encoder, latent space and decoder structure, while the classifier consumes data from the features captured by variational autoencoder in the latent space.

B. Data Selection

In the process of data preprocessing, we obtained audio samples from the Ravdess datasets and custom data from distinct sources. Our primary objective was to construct a coherent emotional audio dataset and convert these audios into Mel spectrograms. Intermediate steps were taken to seamlessly

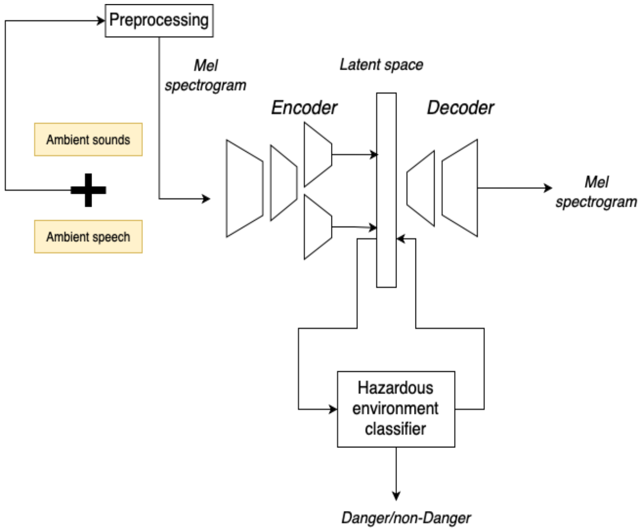

Fig. 1. Proposed architecture of the hazardous classifier model.

integrate these diverse data types, involving tasks such as aligning sampling rates, set at 44.1KHz.

Within the Ravdess dataset, audio clips exceeded three seconds, with approximately 2 seconds of null data. Consequently, trimming zero-data was imperative to obtain a high-quality signal for further processing. Normalization of voices to a standardized volume and noise reduction were performed, particularly essential for downloaded patterns that exhibited varying durations and significant zero data within different frames.

The generation of Mel spectrograms involved utilizing the short time Fourier transform (STFT) technique, followed by mapping the spectrogram to a Mel scale. This method, preconfigured with 128 Mel values, enabled a precise characterization of the audio data.

Emotion selection involved three emotions from the Ravdess dataset, excluding “Disgust” as it does not pertain to a dangerous or non-dangerous environment precisely. The chosen emotions (Neutral, Calm, Happy, Sad, Angry, Fearful) were carefully balanced to ensure equal representation in the dataset.

Regarding ambient noises, glass breaking and gun firing were chosen to represent dangerous environments, while cooking, eating noises, human steps, and opening/closing windows represented safe environments. The specific quantities used are detailed in the experiment section.

In our training approach, we focused on scenarios that involve neutral, calm, and happy emotions coupled with typical indoor noises like cooking, eating, and human movements. The decision to exclude scenarios where both dangerous and non-dangerous noises and speech coexist was deliberate. Training artificial neural networks with such mixed data might lead to the erroneous understanding that during routine, calm, or happy speech, potentially dangerous events like gunfire or breaking glass should be expected. This contradicts real-life situations where such consistency is infrequent and therefore was not incorporated into our training data to maintain the model’s adherence to realistic scenarios.

C. Prosodic and Phonetic Regulariser Features’s Description

Detecting hazardous environments using ambient sounds and speech poses a significant challenge, benefitting greatly from a multifaceted approach involving various audio features. In our research, we focus on employing spectral bandwidth, spectral contrast, and formants 1 to 5 for this purpose.

Formants in speech processing refer to the resonant frequencies of the vocal tract, manifesting as peaks in the sound spectrum. They play a vital role in speech production and perception, representing specific vocal tract configurations through their frequencies. Notably, the first few formants (such as F1, F2, F3, etc.) are pivotal in speech recognition, differentiating speech sounds based on their positions and transitions. Although they are not typically regarded as prosodic features, formants are instrumental in recognizing vowels and consonants, providing essential phonetic information in speech analysis [31].

Spectral contrast, another key feature, quantifies the amplitude disparity between peaks and valleys in the sound spectrum. This metric captures variations in spectral energy, indicating the sharpness or smoothness of transitions between different frequency bands. Drastic changes in spectral contrast signify specific events or objects in the environment. In danger detection scenarios, abrupt increases in spectral contrast can indicate events like glass breaking or gunshots. Monitoring these shifts enables the identification of unusual and potentially perilous situations [32].

Spectral bandwidth, the third feature under consideration, refers to the width of the frequency spectrum of a sound. It measures the dispersion of frequencies around the central frequency. Sounds with broader spectral bandwidths encompass a wider frequency range and are generally classified as broadband sounds. In contexts where danger needs to be identified, wide spectral bandwidths indicate loud and potentially hazardous noises, especially in otherwise quiet settings. For instance, while the sounds of everyday activities like cooking or eating are typically narrowband, noises such as gunfire or explosions produce broad-spectrum signals. Analyzing spectral bandwidth helps recognize the presence of such broad-spectrum events, aiding in the detection of potential threats [32].

By comprehensively analyzing formants, spectral contrast, and spectral bandwidth, an acoustic system can effectively differentiate between normal activities and events that might pose a danger within a specific environment. These distinctive features allow the system to identify specific sound patterns associated with perilous situations, making them invaluable tools in acoustic surveillance and safety systems.

D. Mathematical Definition of Proposed Model’s Regularization

The additional phonetic-prosodic regularisation term $R(phopro(x), phopro(p_\theta(x|z)))$ was added to the well known evidence lower bound (ELBO) of VAE [33], in order to make each input datum x, to remain close to the corresponding decoded datum in the vector of danger $phopro(p_\theta(x|z))$ and $phopro(x)$, respectively.

The danger term to be added is as follows:

$$PhoProDiff\_R\_Loss = \alpha \cdot R(phopro(x), phopro(p_\theta(x|z))) \tag{1}$$

PhoProDiff stands for phonetic-prosodic difference regularization loss.

The prosodic regularised variational autoencoder loss function will finally be defined as follows.

$$\begin{aligned} L(\phi, \theta, x) = E_{q_\theta}(z|x)[log_{p_\theta}(x|z)] - D_{KL}(q_\phi(z|x) \| p(z)) \\ +\alpha \cdot R(phopro(x), phopro(p_\theta(x|z)) \end{aligned} \tag{2}$$

Letting α being $0 \leq \alpha \leq 1$. Assuming E as the expected value, D_{KL} as Kullback-Leibler Divergence, and R as regularisation.

The $R(phopro(x), p_\theta(x|z))$ term is defined as a mean squared error for each spectral feature. Phonetic-Prosodic regulariser over one spectral feature calculation can be defined as follows.

$$\begin{aligned} R(phopro(x), phopro(p_\theta(x|z))) = (phopro(x) \\ -phopro(p_\theta(x|z))^2 \end{aligned} \tag{3}$$

The combination of formants, spectral contrast, and spectral bandwidth extracted from the actual pre-processed Mel spectrogram data is represented as the phonetic and prosodic vector $phopro(x)$. This vector serves the purpose of enforcing regularization, ensuring alignment with the phonetic and prosodic attributes of the speech data.

The regularization based on phonetic and prosodic qualities is applied during the training process [34].

III. Experiments

A. Experiment Details

In our proposed methodology, we devised two distinct models: the Variational Auto-Encoder (VAE) and the danger classifier, each fulfilling specific roles. The proposed VAE operates as an unsupervised learning tool, generating a condensed data representation suitable for tasks like data generation and denoising. However, it is not optimized for direct danger classification.

Conversely, the danger classifier specialises in precisely this task, classifying danger based on the acquired features. It takes encoded features from the danger encoder and associates them with corresponding danger and non-danger classes. This separation allows for independent training processes and facilitates the exploration of various classifier architectures without impacting the proposed VAE. This design ensures the versatility of the learned representation from the proposed VAE for diverse downstream tasks, including danger classification.

Essentially, the proposed VAE learns a meaningful latent representation of input data, which the danger classifier utilizes for classification. This clear division of roles enhances modularity and adaptability in the overall learning process. Our danger classifier follows a supervised learning paradigm, categorizing input data into distinct danger and non-danger classes. Using an emotion dataset containing Mel spectrogram images and corresponding danger labels, the classifier learns to map these spectrograms to specific danger labels.

Our prosodic regularized variational auto-encoder model is trained with emotionally expressive speech audio. We have innovatively incorporated adjustments between speech and ambient noise sounds, introducing a novel approach. Notably, phonetic and prosodic adjustments have never been applied to this kind of input data within an adapted auto-encoder. Implementing our model under these conditions enables us to capture danger data more realistically, emphasizing the value of a generative model in extending real speech with authentic sounds often present in genuine danger environments.

Regarding ambient noises, we utilized a total of 40 audio clips. Ten audio clips were dedicated to glass breaking and gunfire, randomly interspersed with angry, fearful, and sad emotional audio clips. Similarly, there were ten audio clips, for cooking or eating noises and indoor movements (such as household steps), randomly distributed with happy, calm, and neutral emotional audio clips. The glass-breaking sounds varied, encompassing scenarios like breaking a window, objects falling and glass being thrown until breaking, as well as handling glasses, considering the potential harm to third parties in the room or the individual handling them. The gunfire sounds included various types of guns such as standard guns, pistols, and rifles. Additionally, we included gunfire from a distance sufficient to be heard from a room in a house.

For each emotion, we collected four neutral audio clips and eight audio clips from each of the other five emotions, per actor. Our dataset comprises a total of 24 actors, ensuring gender balance. In summary, from each actor, we utilized 40 audio clips, resulting in a total of 1056 audio clips used for training and testing. In our study, we assumed our data was initially separated, and we organized it placing speech at the beginning followed by danger noises, creating 1-second audio segments. While it is ideal for the data to be pre-separated, we consider this task accomplished within our work.

In our speech processing experiments, we utilized two categories of training data: acted speech and real daily conversation speech nuances. Acted speech, found in audiobooks, involves actors simulating emotions. In contrast, daily conversation speech nuances captures natural expressions from sources like YouTube talk-shows, street conversations, and shop dialogues. Both types of data were included in our study, restricted to indoor nuances and speech. Acted speech for testing purposes was sourced from the RAVDESS database [35], while daily conversation nuances data was collected from diverse real-world environment downloaded from Freesound public open datasets.

Our combined dataset merged the RAVDESS database, consisting of facial and vocal expressions in North American English from 24 gender-balanced actors, with custom data containing emotion- and ambient noise in present in the environment. The dataset we compiled included 1056 audio clips used for training and testing, recorded at 44.1kHz, from 12 actors, covering 6 selected emotions out of 8 available emotions. Each emotion was associated with specific patterns, enhancing authenticity. Sentences from the database, such as ”Kids are talking by the door” and ”Dogs are sitting by the door,” were utilized. The training and testing data were divided into 80 percent and 20 percent, respectively.

To maintain dataset consistency, we linked emotions to the primary dataset. For instance, selecting a ”happy” emotion from a male actor involved aligning emotional level sentences with non dangerous ambient nuances, leveraging the similar vocal characteristics in emotional patterns. Ambient noises were randomly chosen while ensuring alignment in events that tend to occur at the same time, or follows to one another, such as angry, sad or scared speakers followed by a glass broken or a gunfire. In the initial tests, 40 audio clips were matched with each corresponding RAVDESS audio danger pair randomly. Importantly, generative models were minimally affected by these variations since they were incorporated during the preprocessing steps.

Both the dangerous environment classifier and the proposed VAE model utilized convolutional layers on pre-processed Mel spectrogram data. The input size for the proposed VAE was 128 by 128 (resized) for both training and testing sets. The proposed VAE’s encoder and decoder consisted of two hidden layers, reducing data dimensions from 128 to 64 and then to 32 in the encoder, and restoring it from 32 to 64 and finally to 128 in the decoder. The proposed VAE featured a single output for encoding and reconstructing data. The emotional classifier received 32-sized data from the proposed VAE encoder and included an output layer with a Softmax activation function corresponding to the six mentioned emotion classes reduced to danger and non-danger opposite classes.

B. Experimental Results

In contrast to traditional methods, our model excels by achieving remarkable results with a limited dataset while capturing intricate patterns present in genuine dangerous environments. Unlike conventional speech models that require extensive datasets for comprehensive testing, our model displays flexibility by leveraging robust, limited nuance patterns present while in the presence of danger. This adaptability ensures precise classification without distorting speaker characteristics or imposing specific positional attributes. Generative models, including our proposed model, comprehend data distributions, enabling classification without excessive reliance on additional patterns.

Nevertheless, our model encounters challenges in generalizing learned sentences across diverse data. However, it excels in recognizing similar sentences and/or noises that share common patterns.

The integration of the phonetic-prosodic regularized VAE model with speech, ambient noises, and the dangerous environment classifier results in enhanced classification accuracy compared to the vanilla VAE with our classifier. Notably, the incorporation of well-defined ambient noises such as gunfire and glass broken like, improves the classification with sad and neutral speech by reducing false positives and negatives in the classification. The proposed VAE adeptly reconstructs patterns collaborating with the classifier, automatically regenerating the input data. The classifier benefit from the latent space features of danger, impacting positively in the classification accuracy.

Our model achieves a test accuracy of 0.924 with extended data, surpassing the vanilla VAE accuracy value of 0.909, and outperforming the standard CNN-based model with 0.742. Furthermore, it delivers superior results in fewer epochs, underscoring its efficiency in accurate emotion classification. However, when compared with the RavdessDB dataset, our model with vanilla VAE with the additional classifier achieves a validation accuracy of 0.575, whereas the proposed VAE with additional classifier achieves 0.56. There are some ambient noises such as the open and close of a windows and the metallic stairs steps that could be confused and further misclassified by AI models, due to their similarity content in other danger noises such as glass broken and long distance gunfire. For human earrings could be perfectly estimated, however for AI models, there is much work to perform and great deals to enhance.

In summary, the adaptability, efficient learning, and enhanced accuracy in environmental hazardous classification make our model a promising advancement in this field.

The training loss/accuracy and the validation loss/accuracy of the vanilla VAE model, can be observed in Fig. 2. At the beginning of the accuracy result image we can observe a jumping until getting a good accuracy of 1, which is not observed in our models for training and validation. The confusion matrices for the training and validation of the vanilla VAE model and the proposed phonetic-prosodic regularized VAE model with our danger classifier are depicted in Fig. 3 and 4, respectively. The classes “0” and “1” represent “Safe” and “Dangerous”, respectively. The training confusion matrix for both models demonstrates accurate predictions for each class. The imbalance in neutral classification is due to the limited number of audios in the neutral class in the Ravdess datasets. During training 287 audios of class “0” (Safe) and 436 audios of class “1” (Danger) are correctly classified, while 97 audios of class “0” and 24 audios of class “1” are misclassifications. For validation, 83 audios of class “0” and 110 audios of class “1” are correctly classified, while 13 audios of class “0” and 6 audios of class “1” are misclassifications. The training loss/accuracy and the validation loss/accuracy of the phonetic-prosodic regularized VAE model with speech and ambient noises, can be observed in Fig. 5.

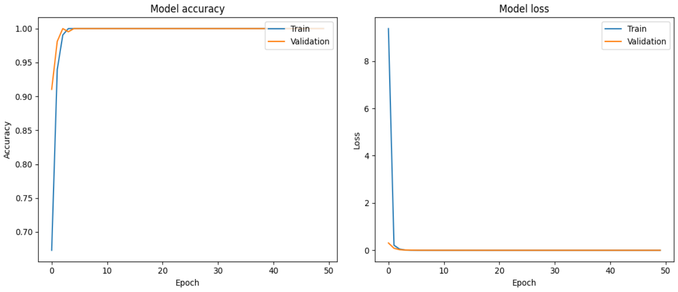

Fig. 2. Training / validation loss and accuracy over epochs of vanilla VAE model.

The confusion matrix for training and validation of the proposed phonetic-prosodic regularized VAE model with speech

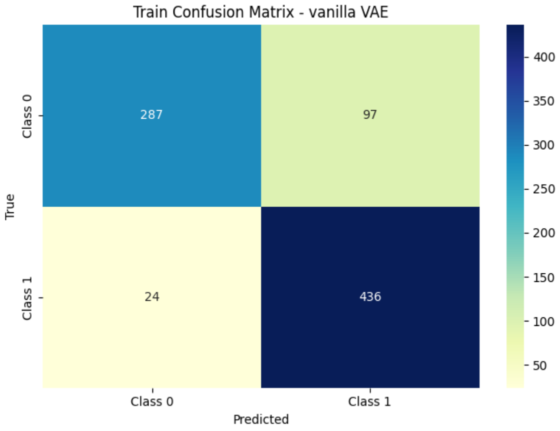

Fig. 3. Confusion matrix for training the vanilla VAE model.

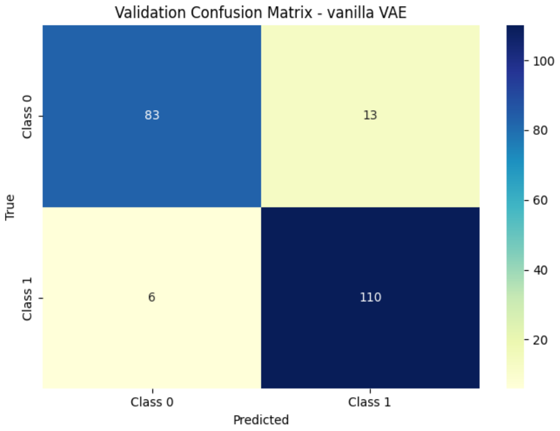

Fig. 4. Confusion matrix for validation of the vanilla VAE model.

and ambient noises, can be observed in Fig. 6 and 7. Phonetic-prosodic regularized VAE model with speech and ambient noises shows better predictions for danger class as “0” and non-danger class as “1”. During training 384 audios of class “0” (Safe) and 460 audios of class “1” (Danger) are correctly classified. For validation, 96 audios of class “0” and 116 audios of class “1” are correctly classified, while there are no misclassifications.

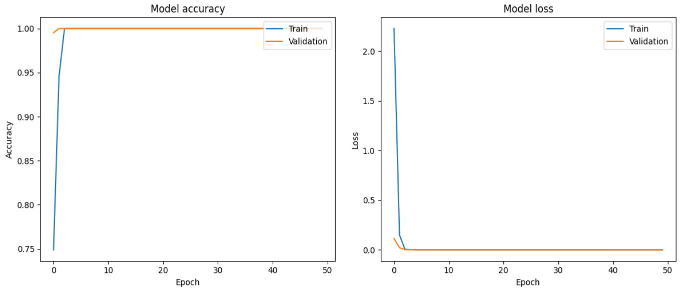

Fig. 5. Training / validation loss and accuracy over epochs of phonetic-prosodic regularized VAE model with speech and ambient noises.

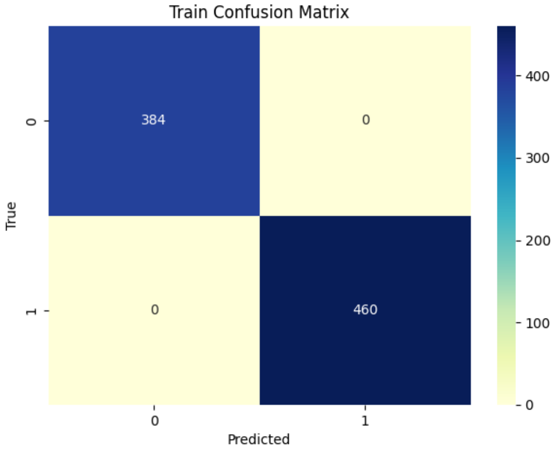

Fig. 6. Confusion matrix for training the phonetic and prosody regularized VAE model with speech and ambient noises.

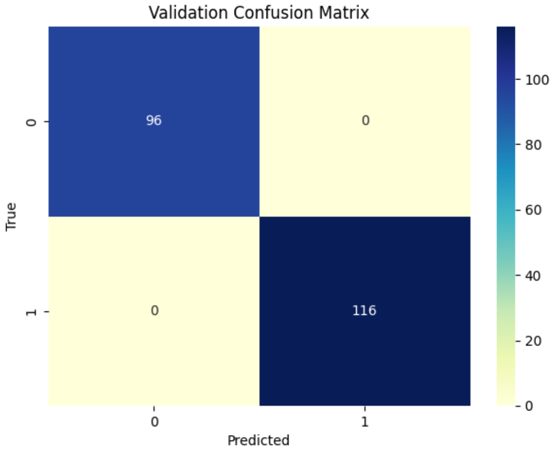

Fig. 7. Confusion matrix for validation of the phonetic and prosody regularized VAE model with speech and ambient noises.

IV. Conclusion

Our proposed model demonstrates notable success in accurate emotion classification, achieving a test accuracy of 0.924 with extended data—an improvement over the vanilla VAE accuracy of 0.909 and the standard CNN-based model’s 0.742. The efficiency of our model is further underscored by its ability to deliver superior results in fewer epochs. However, when confronted with the RavdessDB dataset, our model’s performance, measured by validation accuracy, shows nuances.

The vanilla VAE with an additional classifier achieves 0.575, while the proposed VAE with an additional classifier achieves 0.56. It is important to note that challenges persist, particularly in discerning ambient noises like the opening and closing of windows and metallic stair steps, which may be prone to confusion and misclassification due to their similarity with other potentially dangerous sounds. The complexity of such audio distinctions poses a challenge for AI models, necessitating ongoing efforts for improvement.

In summary, our study introduces a method leveraging generative models, specifically a variational autoencoder, for identifying hazardous environments through the analysis of emotional speech and ambient noises.

Our model, integrating phonetic and prosody features, addresses disparities between input and expected data. As part of our future research, we aim to explore areas such as background noise analysis, the separation of speech and ambient sounds, and the potential extension of our work to real-time danger processing and analysis.

References

[1] Smailnov, Nurzhigit, et al. A Novel Deep CNN-RNN Approach for Real-time Impulsive Sound Detection to Detect Dangerous Events. International Journal of Advanced Computer Science and Applications, vol. 14, no 4, 2023.

[2] Ribino, Patrizia; Lodato, Carmelo. ”A distributed fuzzy system for dangerous events real-time alerting”. Journal of Ambient Intelligence and Humanized Computing, vol. 10, p. 4263-4282, 2019.

[3] Carbonneau, Marc-André, et al. ”Detection of alarms and warning signals on a digital in-ear device”. International Journal of Industrial Ergonomics, vol. 43, no 6, p. 503-511, 2013.

[4] Wang, Qing, et al. A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, p. 1251-1264, 2023.

[5] Wang, Qing, et al. The NERC-SLIP system for sound event localization and detection of DCASE2022 challenge. DCASE2022 Challenge, Tech. Rep., 2022.

[6] Sharath Adavanne, Archontis Politis et al. Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks.https://doi.org/10.48550/arXiv.1807.00129, Dec. 17, 2018.

[7] K. Lopatka, J. Kotus et al. ”Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations”. Multimed. Tools Appl.,DOI 10.1007/s11042-015-3105-4, vol. 75, pp.10407-10439, 2016.

[8] Brian Gygi, Valeriy Shafiro; Environmental sound research as it stands today. Proc. Mtgs. Acoust. https://doi.org/10.1121/1.2917563, vol. 1, no. 1, June 2007.

[9] Toni Heittola, Annamaria Mesaros, Tuomas Virtanen, and Moncef Gabbouj, “Supervised model training for overlapping sound events based on unsupervised source separation,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8677–8681, 2013.

[10] Selina Chu, Shrikanth Narayanan, and CC Jay Kuo, “Environmental sound recognition with time–frequency audio features,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1142-1158, 2009.

[11] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen, “Context-dependent sound event detection,” EURASIP Journal on Audio, Speech, and Music Processing, vol. 2013, no. 1, pp. 1-13, 2013.

[12] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen, “Acoustic event detection in real life recordings,” in 18th European Signal Processing Conference, pp. 1267-1271, 2010.

[13] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Vir- tanen, “Polyphonic sound event detection using multi label deep neural networks,” in IEEE International Joint Conference on Neural Networks (IJCNN), 2015.

[14] Elelu, Kehinde; LE, Tuyen; LE, Chau. Collision Hazard Detection for Construction Worker Safety Using Audio Surveillance. Journal of Construction Engineering and Management, vol. 149, no. 1, 2023.

[15] Svatos, Jakub; HOLUB, Jan. Impulse Acoustic Event Detection, Classification, and Localization System. IEEE Transactions on Instrumentation and Measurement, vol. 72, p. 1-15, 2023.

[16] A. Ilic Mezza, G. Zanetti, M. Cobos and F. Antonacci, ”Zero-Shot Anomalous Sound Detection in Domestic Environments Using Large-Scale Pretrained Audio Pattern Recognition Models,” ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, doi: 10.1109/ICASSP49357.2023.10095736., pp. 1-5, 2023.

[17] K. Lopatka, J. Kotus, A. Czyzewski1, ”Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations”, Multimed. Tools Appl., 75, pages 10407-10439, Dec. 2016.

[18] K. Zhou, B. Sisman, et Al., ”Vaw-gan for the disentanglement and recomposition of emotional elements in speech”, https://doi.org/10.48550/arXiv.2011.02314, Nov., 2020.

[19] A. H. Liu, et Al, ”Towards unsupervised speech recognition and synthesis with quantized speech representation learning”, in IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), IEEE, pp. 7259-7263, 2020.

[20] Y. Gao, R. Singh, et Al, ”Voice impersonation using generative adversarial networks”, in IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), IEEE, pp. 2506-2510, 2018.

[21] S. Leglaive, L. Girin, and R. Horaud, ”Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization”, in Proc. of ICASSP, Brighton, UK, 2019.

[22] X. Li, M. Akagi, ”A three-layer perception model for valence and arousal-based detection from multilingual speech”, in Proc. Interspeech 2018, pp. 3643-3647, 2018.

[23] U. Tiwari, M. Soni, R. Chakraborty, A. Panda, and S. K. Kop- parapu, “Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7194-7198, 2020.

[24] J. Zhao, S. Chen, ”Speech emotion recognition in dyadic dialogues with attentive interaction modeling”, in Proc. Interspeech 2019, pp. 1671-1675, 2019.

[25] J. Liu et Al., ”Temporal attention convolutional network for speech emotion recognition with latent representation”, in Proc. Interspeech 2020, pp. 2337-2341, 2020.

[26] K. Akuzawa, Y. Iwasawa, et Al., “Expressive speech synthesis via modeling expressions with variational autoencoder,” in Proc. of Interspeech, Hyderabad, India, 2018.

[27] Xue Feng, Yaodong Zhang, and James Glass, “Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition,” in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, pp. 1759–1763, 2014.

[28] M. Blaauw, J. Bonada, ”Modeling and transforming speech using variational autoencoders”, in Proc. Interspeech 2016, pp. 1770-1774, 2016.

[29] C. H. Wu, et Al, ”Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis”, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1394-1405, 2009.

[30] Y. Xu, “Speech prosody: A methodological review,” Journal of Speech Sciences, vol. 1, no. 1, pp. 85-115, 2011.

[31] G. Zhang, S. Qiu, et Al., ”Estimating mutual information in prosody representation for emotional prosody transfer in speech synthesis”, in Proc. of ISCSLP, pp. 1-5, 2021.

[32] Latif, Siddique and Rana, Rajib and Qadir, Junaid and Epps, Julien. ”Variational autoencoders for Learning latent representations of speech emotion,” December 2017.

[33] D. P.. Kingma, M. Welling, ”Auto-encoding variational Bayes”, in Proc. 2nd International Conference on Learning Representations, 2014.

[34] R. A. Khalil, et Al., ”Speech emotion recognition using deep learning techniques: A review”, in IEEE Open Access journal, doi 10.1109/ACCESS.2019.2936124, vol. 7, pp. 117327-117345, 2019.

[35] Livingstone SR, Russo FA. 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, doi: 10.1371/journal.pone.0196391, vol. 13, no. 5, May, 2018.