ADScAI
Applied Data Science & Artificial Intelligence Symposium

Cross-Domain Bimodal SER for Customer Service and TV Show Domains

1st Naethree Premnath
Department of Statistics
University of Colombo
Colombo, Sri Lanka
0009-0008-7431-7023

2nd Pemantha Lakraj
Department of Statistics
University of Colombo
Colombo, Sri Lanka
0000-0003-3921-8552

3rd Yasas Jayaweera
Department of Statistics
University of Colombo
Colombo, Sri Lanka
0009-0006-0532-7143

Keywords—Cross-domain SER, Multimodal SER, Customer Service, TV Show, Context

I. INTRODUCTION

Human speech is the most common and expedient way of communication, and understanding speech is one of the complex mechanisms that the human brain performs. As technology advances, replicating this ability in machines has become essential, leading to the rise of Speech Emotion Recognition (SER) as a key field in artificial intelligence and human-computer interaction. However, the challenge of accurately recognizing emotions from speech is compounded by the variability in emotional expression across different contexts [1]. In customer service interactions, emotions like happiness or frustration are often conveyed subtly, whereas in TV shows, they are exaggerated for dramatic effect. This contrast poses a challenge for SER models, as emotional expressions differ significantly across domains.

A. Background

With the evolution of SER, researchers have progressively tackled challenges related to dataset variability. Early studies focused on feature normalization to address dataset differences [2], while later work explored domain adaptation to improve model transferability [3]. However, these approaches primarily address cross-corpus variability, such as recording conditions or linguistic differences, but do not consider the influence of domain information and context. This gap persists despite evidence that context critically shapes emotion perception [4]. For instance, frustration in customer calls differs acoustically and semantically from scripted anger in TV shows, yet existing studies rarely compare cross-domain performance.

Domain-specific emotion recognition in speech, focusing solely on call center applications, has been explored [5], but these studies lack comparisons across domains, limiting generalizability. Moreover, their datasets are not publicly available, preventing further research and model improvements. Even when studies have researched different contexts, only a few have focused on studying the impact of context on system accuracy [6]. Compounding this, the field's reliance on scripted datasets with exaggerated emotions and the scarcity of real-world data severely limits real-world applicability [7]. Urgent efforts are needed to develop context-aware models and diverse datasets for applications like customer service and healthcare.

Furthermore, multimodal SER improves accuracy by combining audio and text to capture both vocal expressions and semantic meaning [8]. While combining lexical cues with acoustic features boosts performance, bimodal approaches have not been evaluated across domains. This study aims to fill that gap by testing bimodal models in both domains.

II. METHODOLOGY

The methodology was structured into three phases: Data Preparation, Domain-Specific Hybrid Model Development and Cross-Domain Evaluation.

A. Data Preparation

Two datasets were utilized: the novel CVEAD (Customer Voice Emotions Analysis Dataset) for customer service interactions and the existing MELD (Multimodal EmotionLines Dataset) for TV show dialogues. CVEAD was built from YouTube mock calls and additional recordings from 12 individuals, annotated via majority voting and transcribed using OpenAI's Whisper model.

Fig. 1 and Fig. 2 show class imbalance in both datasets, addressed using four strategies: SMOTE, undersampling, augmentation, and augmentation with SMOTE. For the latter, 256 combinations of augmentation counts (0, 20, 40, 60) were tested. The best strategy was selected based on the cross-validated F1 score with a Random Forest classifier. Audio augmentation included Gaussian noise, time stretching, and pitch shifting, while text augmentation used back translation, synonym replacement, and random word insertion. Feature extraction involved 41 features for audio (MFCCs, Chroma, Spectral features, Tonnetz, Zero Crossing Rate) and the top 1000 TF-IDF vectors for text.

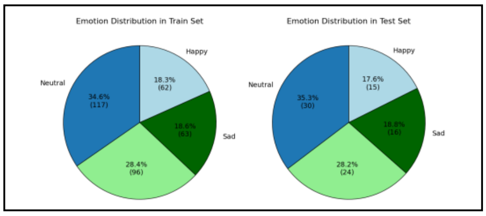

Fig. 1. Emotion Distribution of Train and Test sets of CVEAD

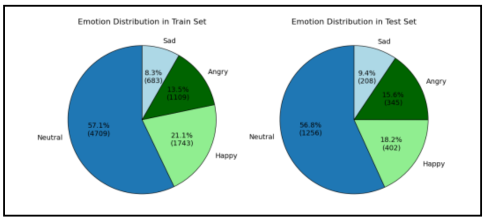

Fig. 2. Emotion Distribution of Train and Test sets of MELD

https://doi.org/10.31705/ADScAI.2025.35

B. Domain-Specific Hybrid Model Development

Machine learning models were trained separately on audio and text modalities. The audio model used 41 features, while the text model used 1000 features. To reduce dimensionality and retain key information, PCA and feature selection methods were applied: Permutation Importance for audio (due to feature complexity and correlation) and Chi-Square for text (effective with TF-IDF vectors). The best strategy was selected based on test set performance. Hybrid models were then created using model-level fusion techniques like Soft Voting, Weighted Voting, and Fuzzy Integration, with the best strategy chosen based on test performance.

C. Cross-Domain Evaluation

To assess the generalizability of the models, cross-domain evaluation was conducted by testing models trained on one dataset (CVEAD or MELD) on the other domain. This helped identify challenges in transferring emotion recognition capabilities across different contexts.

III. RESULTS AND DISCUSSION

For the best class balancing strategy, in CVEAD audio, the most effective approach was augmentation followed by SMOTE, with the optimal augmentation count being 0 for Angry, Happy, and Sad emotions, and 60 for Neutral. CVEAD text achieved the best results with augmentation alone. For MELD, undersampling was preferred for both audio and text to avoid excessive artificial data generation due to severe class imbalance.

The optimal dimensionality reduction and feature selection strategies were chosen based on their ability to retain key information while improving model performance. For CVEAD audio, Permutation Importance with feature correlation removal resulted in 15 features that best captured relevant audio characteristics. PCA reduced CVEAD text to 259 components, balancing information retention with complexity. In MELD, PCA retained 14 components for audio, offering the best trade-off between complexity and performance, while Chi-Square feature selection for text retained 137 features that best captured relevant terms.

Table I summarizes the best models, while Table II presents the cross-evaluation results of hybrid models fused by Weighted Voting, which achieved the highest F1 scores. Audio and Text weights are also shown in Table II.

TABLE I. RESULTS OF BEST MODELS

	Best Model	*F1 Score*
CVEAD Audio	KNN (Metric=euclidean, No. of Neighbors=3, Weights=distance)	75.37%
CVEAD Text	SVM (C=10, Gamma=scale, Kernel=rbf)	66.67%
MELD Audio	Logistic Regression (C=10, Penalty: l2, Solver: saga)	38.24%
MELD Text	SVM (C=10, Gamma=scale, Kernel=linear)	51.70%

TABLE II. RESULTS AFTER CROSS-DOMAIN EVALUATION

	F1 score on CVEAD test set	*F1 score on MELD test set*
CVEAD Hybrid Model (Audio: 0.45, Text: 0.55)	80.97%	43.36%
MELD Hybrid Model (Audio: 0.25, Text: 0.75)	27.56%	52.09%

Table II highlights the challenges in adapting emotion recognition models across domains. Models trained on customer service data performed well within their domain but saw a nearly 50% drop in performance on MELD data, and vice versa. Analysis of audio features revealed MFCCs as crucial in both domains, reflecting their importance in capturing emotional cues. However, feature significance varied: Spectral Rolloff was more relevant in MELD, reflecting TV dialogue dynamics, while it was less prominent in CVEAD's controlled customer service speech. Textual analysis showed domain-specific differences, with words like "issue" and "refund" in CVEAD reflecting task-oriented interactions, and informal terms like "aww" and "damn" in MELD capturing emotional expressions. These findings emphasize the need for domain-specific models and the importance of domain-specific feature engineering.

IV. CONCLUSION

This study examines cross-domain SER challenges in customer service and TV dialogues, revealing domain-specific traits that hinder generalization. Customer service data features subtle, task-oriented emotions, while TV shows exhibit exaggerated expressions, leading to performance drops. These findings have key implications for customer service and media analysis. In customer service, recognizing subtle emotions improves automation and satisfaction. In media analysis, understanding emotional dynamics enhances content recommendation and viewer engagement.

V. FUTURE WORK

However, limitations such as CVEAD's size should be addressed by expanding the dataset, which would enhance SER for customer service. Future work should also explore additional domains like healthcare and emergency call centers. Integrating visual features could further improve the SER model and pave the way for more valuable insights.

REFERENCES

[1] A. Tawari and M. M. Trivedi, "Speech emotion analysis: Exploring the role of context," IEEE Trans. Multimedia, vol. 12, no. 6, pp. 502–509, Oct. 2010, doi: 10.1109/TMM.2010.2058095.

[2] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," IEEE Trans. Affect. Comput., vol. 1, no. 2, pp. 119–131, 2010, doi: 10.1109/T-AFFC.2010.8.

[3] N. Liu, Y. Zong, B. Zhang, L. Liu, J. Chen, G. Zhao, and J. Zhu, "Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Apr. 2018, pp. 5144–5148, doi: 10.1109/ICASSP.2018.8461848.

[4] U. Hess and S. Hareli, "The impact of context on the perception of emotions," in The Expression of Emotion: Philosophical, Psychological and Legal Perspectives, pp. 199–218, 2016. doi: 10.1017/CBO9781316275672.010.

[5] B. Waelbers, S. Bromuri, and A. P. Henkel, "Comparing neural networks for speech emotion recognition in customer service interactions," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2022, doi: 10.1109/IJCNN55064.2022.9892165.

[6] Chenchah, F., & Lachiri, Z. (2014). Speech Emotion Recognition in Acted and Spontaneous Context. Procedia Computer Science, 39(C), 139–145. https://doi.org/10.1016/J.PROCS.2014.11.020.

[7] R. Pereira et al., "Systematic review of emotion detection with computer vision and deep learning," Sensors (Basel, Switzerland), vol. 24, no. 11, 2024. doi: 10.3390/S24113484.

[8] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in Proc. 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 112–118. doi: 10.1109/SLT.2018.8639583.