IDEAS EN CIENCIAS DE LA INGENIERÍA

Exploración de los Límites del Muestreo Selectivo para el Reconocimiento de Emociones del Habla Dependientes del Hablante

Exploring Selective Sampling Bounds for Speaker-Dependent Speech Emotion Recognition

Castorena, C.[1]*, Lopez-Ballester, J.[1], Ferri, F.J.[1], Cobos, M.[1],
(*carlos.castorena@uv.es)

Recibido: 2024-11-12 Aceptado: 2024-11-27

Resumen

El reconocimiento de emociones en el habla es crucial para diversas aplicaciones, desde el ámbito de la salud mental hasta la interacción con máquinas. Aunque los modelos basados en aprendizaje profundo han demostrado ser efectivos en este tipo de tareas, su rendimiento puede verse afectado cuando se enfrentan a nuevos individuos, incluso si se entrenaron con grandes y diversos conjuntos de datos. En este estudio, se analizó la relevancia de ajustar y personalizar estos modelos para mejorar la precisión en el reconocimiento emocional. A pesar de que los modelos generales ofrecen buenos resultados iniciales, se percibe que su capacidad para predecir emociones disminuye cuando se aplican a nuevos sujetos. A través de este trabajo, se muestra cómo personalizar un modelo con una pequeña cantidad de muestras de un nuevo individuo puede aumentar considerablemente su desempeño, utilizando técnicas como el ajuste fino y la transferencia de aprendizaje. Finalmente, se discuten las implicaciones de estos resultados en aplicaciones reales y se. sugieren posibles direcciones para futuras investigaciones en la detección de emociones mediante inteligencia artificial.

Palabras clave: reconocimiento de emociones, deep learning, adaptación al dominio.

Abstract

Emotion recognition in speech is crucial for various applications, ranging from mental health to human-machine interaction. Although deep learning-based models have proven effective in these tasks, their performance can be affected when faced with new individuals, even if they were trained with large and diverse datasets. In this study analyze the importance of fine-tuning and personalizing these models to improve emotional recognition accuracy. Despite the good initial results of general models, we

[1]Departament d'Informàtica, Universitat de València, Burjassot, Spain. 46100.

observe a decline in their predictive ability when applied to new subjects. Through this work, it is demonstrated how personalizing a model with a small number of samples from a new individual can significantly improve its performance, using techniques such as fine-tuning and transfer learning. Finally, the implications of these findings in real-world applications are discussed, and possible directions for future research in emotion detection through artificial intelligence are suggested.

Keywords: emotion recognition, deep learning, domain adaptation.

INTRODUCTION

Speech Emotion Recognition (SER) is a significant challenge in natural language processing and human-machine interaction, as it allows the identification of emotions based on acoustic signals. This technology has practical applications in various fields: in customer service, where it monitors the user's emotional state to improve interaction quality by detecting frustration or dissatisfaction [1]; in healthcare, where SER facilitates mental health monitoring by providing essential information for diagnosing and treating emotions such as sadness or fear [2]; and in the automotive industry, where it can be used to assess the driver's emotional state, enhancing both safety and comfort.

Although current systems efficiently process speech for tasks such as transcription or automatic translation [3], SER faces significant limitations. One key challenge lies in the subjective perception of emotions by annotators [4], a complex issue as different people may interpret varying emotional content from the same audio. Furthermore, the diversity of speakers, contexts, and environments impacts model robustness. Additionally, SER is influenced not only by the acoustic characteristics of the voice but also by speech content; while these semantic associations can improve performance, they may also introduce challenges as they are not reliably transferable to new contexts [5].

Emotions in SER can be represented through either categorical or dimensional labels, each suited to different aspects of emotion recognition. Categorical representations label emotions as discrete classes (such as happiness, anger, sadness, and neutrality) making classification straightforward and widely interpretable. This approach is often advantageous, as it simplifies both model design and performance evaluation by aligning with commonly understood emotional states. Moreover, categorical labels can be more robust and consistent across different speakers, languages, and cultural contexts, which is especially beneficial in practical applications requiring clear emotion identification.

Dimensional representations, in contrast, define emotions by continuous variables, typically arousal (intensity), valence (positivity/negativity), and dominance (control), allowing a nuanced analysis of emotional states. While these continuous labels offer flexibility to capture subtle variations within each emotion, they add complexity to model training and interpretation. The dimensional approach also requires meticulous labeling, as subjective interpretations of arousal, valence, and dominance may vary significantly among annotators. As a result, categorical representations generally offer a practical balance of simplicity and clarity, making them well-suited for many SER tasks, whereas dimensional representations may be preferred for specialized applications that demand fine-grained emotional assessment.

In traditional SER, when a model is trained and evaluated with the same speaker, it is known as speaker-dependent SER [6]. These models tend to achieve high levels of accuracy [7, 8]. However, their applicability is limited due to low generalization capacity. By contrast, speaker-independent SER

https://doi.org/10.36677/rici.v3i1.25132 Enero-Julio 2025 ISSN: 2992-7447 págs. 48-58

[6] presents a greater challenge, as it requires the model to recognize emotions in a generalized way across individuals, each with unique emotional expressions and acoustic characteristics. This demands that models adapt to effectively handle variations not fully represented in the training phase.

To address these limitations, speaker-independent SER studies often rely on domain adaptation techniques, where the source data representation (training speakers) is adjusted to resemble the characteristics of new speakers. For example, Lu et al. [9] focused their research on measuring domain shift between source and target data to reduce this discrepancy. AbdelWaha et al. [10, 6, 4] have made significant contributions to this field: in 2015, they presented a method to adapt the target domain using different percentages of target domain samples; in 2017, they introduced an active learning-based method to incorporate these samples into source domain training, combined with domain adaptation strategies; and in 2018, they proposed the use of adversarial multitask learning to align representations between training and test domains using a gradient reversal layer. These studies aim to develop models capable of generalizing with a relatively low number of new samples.

In this work, we explore how, through the use of a small set of labeled samples, it is possible to fine-tune a pre-trained model to improve its accuracy in recognizing emotions in a new speaker. We first analyze the model's performance using a random selection of samples, then compare its performance when using "key samples," which have a significant impact on classification compared to others. Although there are previous studies along this line [10, 11], most do not fully explore classifier capacities when working with a small number of samples, limiting themselves to variations in training set size without examining optimal sample selection. The structure of this work is organized as follows: in the Methodology section, we describe the data, the model used, and the strategy applied for adapting to new speakers. In the Results and Discussion section, we analyze and discuss the main findings obtained from our experiments. Finally, in the Conclusions and Future Work section, we summarize key insights and propose directions for future research.

Methodology

Dataset

The landscape of available datasets for emotion detection in speech is diverse, with each resource distinguished by unique attributes, such as the number of participants, represented languages, types and quantities of emotions, and sample count. For this study, we drew on multiple datasets that are commonly applied in SER research, each contributing specific advantages to the robustness of our model. All datasets consist of audio samples with a sampling rate of 16,000 Hz and durations ranging from 15 to 60 seconds. Each audio recording features isolated speech from a single speaker without background noise, and every audio sample is assigned a single, unique emotion label.

The MSP-Podcast dataset [12] offers a substantial foundation with its 83 hours of English podcast recordings, annotated for valence, activation, and dominance dimensions. This corpus includes contributions from 60 speakers (32 male and 28 female), with sample counts per speaker ranging widely from 42 to 912, resulting in a total of over 50,000 samples. This dataset's dimensional labeling provides a nuanced approach to capturing emotional expression in naturalistic settings. Similarly, the IEMOCAP dataset [13] offers 12 hours of dialog interactions, including both scripted and improvised segments from 10 speakers (balanced with 5 men and 5 women). The dataset provides 10,039 labeled samples, allowing

for a comprehensive view of emotional dynamics in conversational speech.

The Berlin Emotional Speech Database (EMO-DB) [14] contributes a contrasting structured approach with recordings from 10 professional actors (5 men and 5 women) enacting seven discrete emotions, yielding a dataset of 800 samples. Each sample is short but precisely labeled, allowing a focused analysis of fundamental emotional expressions. From a Spanish-language perspective, the Mexican Emotional Speech Database (MESD) [15] broadens our linguistic scope, featuring recordings from three speakers (a man, a woman, and a child) across six emotions with 48 samples per emotion and speaker, totaling 864 samples.

For additional variation in emotional expression, the RAVDESS dataset [16] includes 7,356 recordings from 24 native English-speaking actors, who express a range of emotions, including calm, happy, sad, angry, fearful, surprise, and disgust. Finally, the Emotional Speech Dataset (ESD) [17] stands out for providing the highest number of samples per emotion per individual, with 350 dialogues for each emotional category. This abundance of samples per speaker makes it ideal for experiments involving new speakers, offering a solid base for adaptation tasks. The dataset includes both native English and Mandarin speakers, with 10 speakers from each language (5 men and 5 women) expressing emotions across five categories: angry, happy, neutral, sad, and surprise.

To maintain consistency in the analysis, we focus on five core emotions (Happiness, Anger, Sadness, Surprise, and Neutral) and work with categorical labels across these datasets, discarding any additional labels. This standardization ensures comparability across datasets.

Base Model

Our approach utilizes a deep classification model pre-trained on data from diverse speakers, as shown in Figure 1. We begin with a modified version of Wav2Vec [5], initially trained on the MSP-Podcast dataset and validated with IEMOCAP, to obtain a deep audio representation. Wav2Vec, a transformer-based architecture, was originally designed to extract both acoustic and semantic features from speech and is primarily used for tasks like automatic speech recognition. Unlike strategies based solely on speech formants [18] or other handcrafted features such as pitch [19], energy [20], and spectral descriptors [21], which focus on specific phonetic or prosodic characteristics, Wav2Vec seeks to capture a broader range of audio features, encompassing both phonetic and semantic information. This comprehensive feature extraction is particularly advantageous for capturing the nuanced and multidimensional nature of emotional expression in speech. Given these capabilities, Wagner et al. adapted Wav2Vec to SER by retraining it specifically for emotion recognition, allowing the model to leverage both acoustic and semantic content for emotion prediction. Notably, their adaptation focuses on dimensional labels, representing emotions within a three-dimensional space of Arousal, Valence, and Dominance.

In our approach, we utilize only the internal representation generated by Wav2Vec, substituting their regression module with our own classification module, which consists of three dense layers with 64, 32, and 5 neurons. The first two layers use ReLU activation functions, while the final layer applies Softmax to produce categorical predictions. To train this classifier module, we use the EMO-DB, MESD, and RAVDESS training partition datasets (70% of the samples), covering a total of 37 speakers. Training is conducted for 200 epochs, with a batch size of 64, using Categorical Cross-Entropy as the loss function. Throughout training, the feature extraction module remains untrainable, focusing all training on the classification module.

https://doi.org/10.36677/rici.v3i1.25132 Enero-Julio 2025 ISSN: 2992-7447 págs. 48-58

This model, referred to as the base model, serves as the foundation for subsequent experiments. Using the final state of this base model, we will apply adaptation techniques to fine-tune the classifier for effective emotion recognition in new speakers.

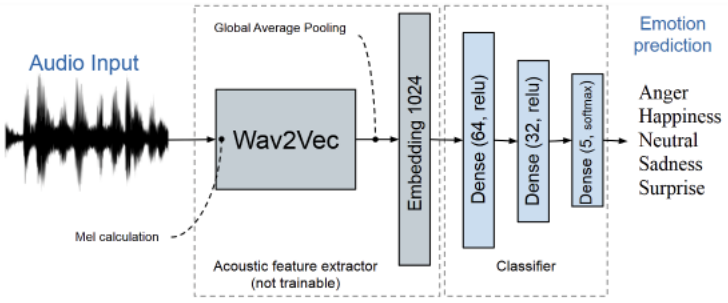


Figure 1: *A pre-trained base model for the classification of 5 emotions.*

Strategies for New Speakers

Our strategy aims to fine-tune the base model's weights to better recognize emotions in new speakers from the ESD dataset. We selected 2 native English speakers and 2 native Mandarin speakers, dividing each speaker's data into separate Training and Testing sets with a 70-30 split. The Training set includes all available samples for potential selection (although not all are necessarily used in each experiment), while the Testing set remains fixed for each speaker throughout all trials.

It is essential to note that each new speaker is processed independently; our approach is designed to improve performance specifically for one new speaker at a time.

The first approach, called Random, incrementally selects 5 random samples (one for each emotion) from the Training set at each step. These samples, combined with those gathered from previous steps, are used to fine-tune the base model over 30 epochs, with only the weights in the classifier layers updated. This incremental process is repeated up to 5 times, resulting in sample sizes ranging from 5 to 25 samples, while using the same loss function configuration as the base model to retrain it. At each step, the model's performance is evaluated on the Testing set using accuracy scores (ACC, percentage of correctly labeled samples), with class-balanced representation across all 5 emotion categories. Each experiment is repeated 5 times per speaker to ensure reliability.

On the other hand, the strategy called Best, selects the most representative samples at each step. At each increment, 200 random sets of 5 samples are evaluated, and the set achieving the highest ACC scores on the Training set is selected. These top-performing samples are accumulated across steps to further refine the model. The base model parameters in the Best strategy remain consistent with those used in the Random approach.

RESULTS AND DISCUSSION

After training, the base model achieved ACC values close to 90% on the test partitions of EMO-DB, MESD, and RAVDESS, as shown in the initial step of Figure 4. It is important to note that these dataset partitions do not ensure speaker independence, meaning the same speaker may be present in both the training and test sets. This result underscores the model's effectiveness in emotion recognition within a speaker-dependent setup, aligning with similar results in recent state-of-the-art studies, and serves as a promising baseline for extending its application to a speaker-independent context.

On the other hand, when the model is tested with new speakers, its performance declines significantly. Figure 2 illustrates the ACC values achieved by two new English-speaking individuals. First, we observe the dashed line marked as "Reference," which indicates the ACC obtained by the base model for the new speaker without any adjustment, yielding scores far below the 90% achieved on the source domain test partition (0.41 and 0.54, respectively). One might assume that this drop in performance could stem not only from speaker domain differences but from other speaker-specific factors. To investigate this further, we calculated the ACC that could be achieved if we were to use all available samples in the training partition for each new speaker; this is marked in the figure by a dashed line labeled "Maximum." In both cases, the results show a significant gap between the Maximum and Reference values, suggesting that the model does indeed have the capacity to classify emotions for these speakers effectively. However, for the first speaker, the ACC value is considerably lower, reaching only 61%, while the second speaker achieves an ACC closer to speaker-dependent results, with 84%. This disparity underscores the challenge models face in generalizing effectively when encountering new speakers, highlighting the complexities of speaker-independent emotion recognition.

The solid black line in Figure 2 shows the average ACC across the five repetitions of the Random selection strategy. In both experiments, we observe that as the model is retrained with a greater number of samples, its performance tends to improve. Notably, after the first step (using only 5 samples), ACC increases significantly compared to the initial state, though subsequent increments yield less pronounced gains. On the other hand, the blue line (Best mean) and violin plots represent the averages and distributions for the Best selection strategy. Initially, when only 5 samples are used, the difference between Random and Best is negligible since both start with a random sample set, reflected in their approximately equal averages. However, with the next step—where the model is retrained with 10 samples (including the 5 most representative ones identified in the previous step and indicated by the dashed blue line and the label Best max)—a marked improvement emerges compared to the Random selection. This finding suggests that not only is it essential to retrain the model with new samples, but careful selection of these samples can lead to significantly better results, approaching those obtained when the entire training partition for the new speaker is utilized. From subsequent steps onward, the ACC values seem to stabilize, indicating a plateau in the gains achieved by adding more samples.

Strictly speaking, the base model was not trained with Mandarin speakers. Although Wav2Vec may contain some information relevant to Mandarin from its pre-training, the model's behavior with new Mandarin speakers aligns with that observed for English speakers, a language highly represented in the datasets. As shown in Figure 3, the classification problem can be solved when the entire training set for the new speaker is used. With the Random selection approach, performance improves as more sam-

https://doi.org/10.36677/rici.v3i1.25132 Enero-Julio 2025 ISSN: 2992-7447 págs. 48-58

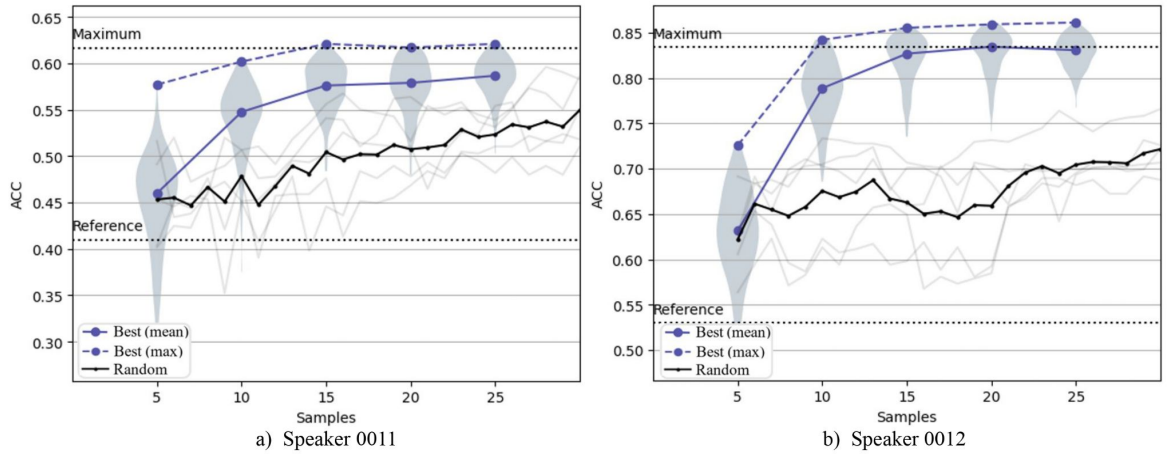


Figure 2: *ACC results for new English speakers with identification codes 0011 and 0012 from the ESD dataset, using a maximum of 25 samples.*

ples are used to retrain the base model; however, it remains consistently below the best strategy, which achieves results equal to or near the Maximum with as few as 10 samples. The model's consistent performance across languages could stem from two factors: first, that the acoustic content and, consequently, the patterns learned by the model may not be language-dependent; or second, that the pre-trained Wav2Vec model retains relevant language information within its deep representations from its unsupervised training phase. In either case, this suggests that language may not be a decisive factor for effective emotion classification.

To analyze the effectiveness of the Random and Best methods, we used the non-parametric Wilcoxon rank-sum test, which is appropriate for comparing the central tendencies of two groups when the samples do not follow a normal distribution or when the sample sizes are unequal, as in this study, where the Random group contains only 5 measures.

The Wilcoxon rank-sum test was applied independently for each speaker at each incremental step to evaluate whether the best method outperforms the Random method. The p-values from these tests can be observed in Table 1. In the first step, when only 5 samples are used, no significant difference is found, as both methods start with a random selection. However, as more samples are added, the p-value tends to decrease, suggesting that the best method becomes increasingly superior to the Random method as the number of processed samples grows.

The accuracy of the same sequence of models in Figure 3, but measured on the test partition of the source dataset is presented in Figure 4. Initially, the model starts with a high ACC of around 90%, as previously discussed. However, as the model is incrementally adjusted with samples from the new speaker, a significant drop in ACC is observed, indicating that the model increasingly prioritizes alignment with the new speaker's characteristics, even at the cost of losing previously learned relationships. After the first few steps, this decline stabilizes, suggesting a local minimum for this particular partition.

The experimental design of this work uncovers several critical points for future contributions in the field. Firstly, while the best strategy demonstrates the effectiveness of intelligently selecting samples,

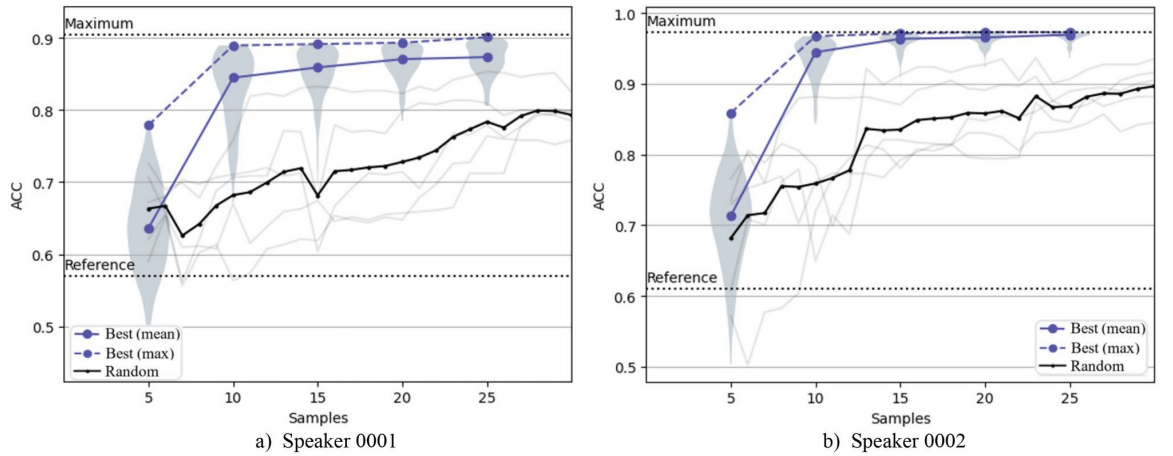


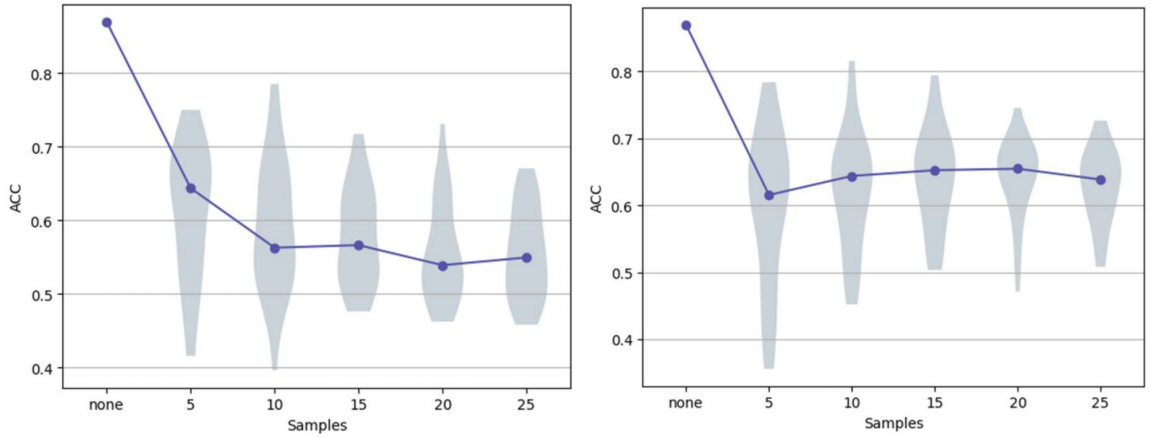
Figure 3: *ACC results for new Mandarin speakers with identification codes 0001 and 0002 from the ESD dataset, using a maximum of 25 samples.*

Table 1: p-values from the Wilcoxon rank-sum tests comparing the performance of the Random and Best methods across incremental steps for each speaker. An asterisk (*) indicates that the null hypothesis (H0) is rejected, and there is a statistically significant difference between the groups at a significance level of α=0.05.

Speaker	5 samples	10 samples	15 samples	20 samples	25 samples
0001	0.0902	0.0001*	0.0001*	0.0001*	0.0030*
0002	0.1700	0.0013*	0.0001*	0.0003*	0.0009*
0011	0.1470	0.0007*	0.0001*	0.0001*	0.0001*
0012	0.1751	0.0021*	0.0002*	0.0001*	0.0001*

https://doi.org/10.36677/rici.v3i1.25132 Enero-Julio 2025 ISSN: 2992-7447 págs. 48-58

it would be impractical in a real-world scenario, as it requires prior knowledge of labels for the new speaker. Nonetheless, it highlights the potential of well-chosen samples to drive significantly better results. Additionally, the evident loss of effectiveness on the original data following retraining underlines the importance of adapting the feature space of the new speaker to be compatible with that of previous speakers. This adjustment will be essential to achieving balanced performance across both the new and original data, paving the way for more robust speaker-independent SER models.



a) Speaker 0001 b) Speaker 0002

Figure 4: *ACC over the source test dataset as the model is retrained using data from new speakers.*

Conclusions

This study has tackled the issue of Speech Emotion Recognition (SER) by customizing a pre-trained model to accommodate new speakers using a limited data set. By employing our approach, which includes both a random and an optimized sample selection strategy, we have shown that the base model's accuracy can be significantly enhanced. Moreover, our results indicate that, even in cases where the optimal samples cannot be precisely identified, a random sampling approach can still lead to meaningful improvements. This insight is especially relevant for real-world applications where comprehensive labeling is impractical. Our findings reinforce the viability of adapting SER models for new speakers with minimal data and underscore the effectiveness of strategic sample selection. This paves the way for further advancements in SER, where models can be tailored more efficiently, requiring fewer resources, thus boosting the potential for enhanced human-machine interaction across diverse applications.

Acknowledgments

We would like to thank the Spanish Research Agency (AEI) and the European Regional Development Fund (ERDF) for partially funding this research through the projects TED2021-131003B-C21 and PID2022-137048OB-C41, funded by MCIN/AEI/10.13039 /501100011033 and the “EU Union NextGenerationEU/PRTR.” Additionally, we thank the Generalitat Valenciana for funding this work through the Santiago Grisolia program (GRISOLIAP/2021/060, CPI-21-232).

DECLARATION OF CONTRIBUTION OF AUTHORS AND COLLABORATORS

Carlos Castorena: Formal analysis, Investigation, Experimentation & Writing. **Maximo Cobos**: Conceptualization, Supervision, Review & Editing. **Jesus Lopez-Ballester**: Review & Editing. **Francesc J. Ferri**: Conceptualization, Supervision, Review & Editing.

REFERENCES

[1] X. Li & R. Lin. Speech emotion recognition for power customer service. *7th International Conference on Computer and Communications (ICCC)*, pp. 514–518, 2021. DOI: 10.1109/ICCC54389.2021.9674619.

[2] N. Elsayed, Z. ElSayed, N. Asadizanjani, M. Ozer, A. Abdelgawad, & M. Bayoumi. Speech emotion recognition using supervised deep recurrent system for mental health monitoring. *IEEE 8th World Forum on Internet of Things (WF-IoT)*, pp. 1–6, 2022. DOI: 10.48550/arXiv.2208.12812.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, & I. Sutskever. Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, pp. 28492–28518, 2023. DOI: 10.48550/arXiv.2212.04356.

[4] M. AbdelWahab & C. Busso. Domain Adversarial for Acoustic Emotion Recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 12, pp. 2423–2435, 2018. DOI: 10.1109/TASLP.2018.2867099.

[5] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, & B. W. Schuller. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 10745–10759, 2023. DOI: 10.48550/arXiv.2203.07378.

[6] M. Abdelwahab & C. Busso. Incremental adaptation using active learning for acoustic emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5160–5164, 2017. DOI: 10.1109/ICASSP.2017.7953140.

[7] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, & S. L. Schneider. Speech emotion recognition using machine learning — A systematic review. *Intelligent Systems with Applications*, vol. 20, 2023. DOI: 10.1016/j.iswa.2023.200266.

[8] A. Hashem, M. Arif, & M. Alghamdi. Speech emotion recognition approaches: A systematic review. *Speech Communication*, vol. 154, 2023. DOI: 10.1016/j.specom.2023.102974.

[9] C. Lu, Y. Zong, W. Zheng, Y. Li, C. Tang, & B. W. Schuller. Domain Invariant Feature Learning for Speaker-Independent Speech Emotion Recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 30, pp. 2217–2230, 2022. DOI: 10.1109/TASLP.2022.3178232.

https://doi.org/10.36677/rici.v3i1.25132 Enero-Julio 2025 ISSN: 2992-7447 págs. 48-58

[10] M. Abdelwahab & C. Busso. Supervised domain adaptation for emotion recognition from speech. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5058–5062, 2015. DOI: 10.1109/ICASSP.2015.7178934.

[11] C. Castorena, J. Lopez-Ballester, M. Cobos & F. J. Ferri. Explorando la personalización de modelos en el reconocimiento de emociones en el habla. *Congreso Ibérico de Acústica – 55º Congreso Español de Acústica – Tecniacústica*, 2024.

[12] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, & S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008. DOI: 10.1007/s10579-008-9076-6.

[13] R. Lotfian & C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, vol. 10, pp. 471–483, 2019. DOI: 10.1109/TAFFC.2017.2736999.

[14] F. Burkhardt, W. F. Sendlmeier, A. Paeschke, M. Rolfes, & B. Weiss. A database of German emotional speech. *Proc. Interspeech*, pp. 1517–1520, 2005. DOI: 10.21437/Interspeech.2005-446.

[15] M. M. Duville, L. M. Alonso-Valerdi, & D. I. Ibarra-Zarate. Mexican emotional speech database based on semantic, frequency, familiarity, concreteness, and cultural shaping of affective prosody. *Data*, vol. 6, p. 130, 2021. DOI: 10.3390/data6120130.

[16] S. R. Livingstone & 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, p. e0196391, 2018. DOI: 10.1371/journal.pone.0196391.

[17] K. Zhou, B. Sisman, R. Liu, & H. Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. *ICASSP*, pp. 920–924, 2021. DOI: 10.48550/arXiv.2010.14794.

[18] Z. T. Liu, A. Rehman, M. Wu, W. H. Cao, & M. Hao. Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence. *Information Sciences*, vol. 563, 2021. DOI: 10.1016/j.ins.2021.02.016.

[19] S. Vaijayanthi & J. Arunnehru. Synthesis Approach for Emotion Recognition from Cepstral and Pitch Coefficients Using Machine Learning. *Lecture Notes in Electrical Engineering*, vol. 733, 2021. DOI: 10.1007/978-981-33-4909-4_39.

[20] R. Thirumuru, K. Gurugubelli, & A. K. Vuppala. Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition. *Digital Signal Processing: A Review Journal*, vol. 120, 2022. DOI: 10.1016/j.dsp.2021.103293.

[21] J. Ancilin & A. Milton. Improved speech emotion recognition with Mel frequency magnitude coefficient. *Applied Acoustics*, vol. 179, 2021. DOI: 10.1016/j.apacoust.2021.108046.