Speech Emotion Recognition for Call Centers using Self-supervised Models: A Complete Pipeline for Industrial Applications

Juan M. Martín-Doñas[1] and **Asier López**[2] and **Mikel de Velasco**[2]
and **Juan C. Vásquez-Correa**[1] and **Aitor Álvarez**[1] and **María I. Torres**[2]
and **Paz Delgado**[3] and **Ane Lazpiur**[3] and **Blanca Romero**[3] and **Irati Alkorta**[4]
[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA)
[2]University of the Basque Country UPV/EHU
[3]NaturalSpeech, [4]Gureak Marketing
jmmartin@vicomtech.org

Abstract

This paper presents a practical methodology to build adapted speech emotion recognition systems in call center scenarios for practical industrial applications. We focus on two specific use cases involving Spanish call centers with different characteristics in order to detect emotional states and improve their protocols. We address all stages of the development process, covering data acquisition, annotation, data harmonization, and model training and evaluation. We rely on cutting-edge self-supervised speech models for classification. This process has been designed to cover an industrial application's needs: data anonymity, reduced costs, and production-level performance. We compare the evaluated methods with well-established research benchmarks to validate our methodology. In addition, a subjective evaluation is performed to analyze their potential use in practical cases. The considered approaches show potential transferable results for these companies in their target call center scenarios.

1 Introduction

Call centers (CC) are increasingly leveraging speech analytics software to automate tasks and extract valuable insights from customer interactions (Hildebrand et al., 2020). This cutting-edge technology analyzes call recordings, enabling companies to enhance their operations. A pivotal aspect of CC conversations involves the speech emotion recognition (SER) of both clients and agents. This paralinguistic information can be used to efficiently transfer a voice call to a physical agent for further queries and discussions, to detect lies or even to identify emotional changes and states (Hema and Marquez, 2023). For instance, promptly identifying frustration can enable agents to employ de-escalation techniques or expedite issue resolution. Moreover, the data obtained can unveil broader trends in customer communication, empowering companies to refine their communication strategies. Therefore, developing reliable SER models holds immense value for the CC market, facilitating deeper customer insights, improved communication strategies, reduced customer frustration, and ultimately, a more positive customer experience (Irastorza and Torres, 2016, 2019).

Recent research on SER has focused on end-to-end deep learning systems, where self-supervised models have shown state-of-the-art (SOTA) performance in common benchmarks (Mohamed et al., 2022). These advancements have primarily been showcased in systems evaluated using acted or elicited databases (Busso et al., 2008). However, the efficacy tends to diminish considerably when these systems are deployed in real-world scenarios characterized by natural speech patterns (Zhu-Zhou et al., 2022). Notably, few works have explored the application of recent advances in SER within real call center environments. The work presented in (Bojanić et al., 2020) is an example of the SER technology application in a prioritizing urgency call system, which was evaluated on an acted Serbian corpus. Deschamps-Berger et al. (2021) evaluated convolutional-recurrent architectures for SER on the French CEMO corpus for medical emergency calls. For the customer service scenario, Pérez-Toro et al. (2021) proposed the classification of emotional states mapped on the arousal-valence dimensions to detect customer satisfaction using acoustic and linguistic models. On the contrary, Parra-Gallego and Orozco-Arroyave (2022) explored the evaluation of prosody and speaker embeddings to detect emotions and customer satisfaction in voicemails. Moreover, Feng and Devillers (2023) studied the continuous SER problem and the use of contextual information during the conversations. Płaza et al. (2022) addressed the database design and the development of solutions for SER classification, focusing on feature extraction methods to model both speech and text data using small classifiers. More recent studies (Deschamps-Berger

et al., 2023; Macary et al., 2023) have proposed using SOTA self-supervised acoustic and language models for SER in CC scenarios, evaluating the performance in French SER corpus for research.

In summary, current technology exhibits considerable potential for enhancing call centers. Nevertheless, its successful implementation in real-world settings needs the resolution of substantial challenges. The establishment of a comprehensive database tailored to specific applications, the creation of precise classification systems or dealing with aspects of conversational speech remain great challenges that demand more research.

This study presents a novel methodology to either build and transfer an adapted SER solution to a real CC application. To this end, we collaborated with two companies providing call center services and designed a customized system for each to meet their specific needs within their respective domains. During the process, we covered the different stages of a suitable system design, starting from the domain data acquisition and annotation, including data pre-processing, manual labeling, and revision of the final corpus. Afterwards, we analysed and tested different feature extraction techniques for classification, from some more traditional to the most recent based on acoustic foundation models, and evaluated several downstream models focused on machine learning and deep learning techniques. In order to deploy practical systems for real scenarios when certain classes are under-represented, we also evaluate the application of binary detection systems that discriminates between neutral and emotional states, showing competitive performance and higher accuracy detection as it is requested in a practical environment.

Our work is complemented by evaluating the proposed solutions on well-established research databases to show that our systems follows the state of the art in the field. Moreover, a subjective comparison between manual and automatic emotional analysis is performed to assess the practical usability in potential real uses cases. Despite the high challenges of the task, our results demonstrate the successful development of transferable SER solutions, addressing the specific needs of the companies involved and holding significant promise for real-world applications.

The rest of the paper is organized as follows. Section 2 analyzes the process for acquiring, processing, and annotating the speech-emotional corpus created in this project. Then, we describe the experimental framework and evaluation results in Section 3. Finally, Section 4 summarizes the conclusions and possible research lines for future work.

2 Building a SER corpus for Call Centers

2.1 Tasks

In this project, we have collaborated with two Spanish companies providing CC solutions to build SER systems adapted to their unique needs and help them improve their interaction protocols and internal quality processes.

Since each company is dedicated to a different market, we tackled the SER problem from two different CC contexts. The first use-case (*CC-Support*) focuses on customer support, where customers often encounter issues like electronic signatures or login problems. The second one (*CC-Debt*) contains phone calls about debt collection, where stronger and more negative emotions usually arise. In this context, customers express frustration and anger more often than in CC-Support calls.

2.2 Data collection and annotation

The construction of the SER systems started with each call center providing approximately 60 hours of valuable speech in-domain data. Due to the calls' nature, containing personal and confidential information, the data could not be labeled in conventional platforms like Amazon Mechanical Turk[1]. Instead, the data were labeled by annotators trained for the task. Consequently, based on the expertise of the annotators and the resources available, we decided to use two annotators per segment instead the conventional three or more (Busso et al., 2008; Parada-Cabaleiro et al., 2018; Vázquez et al., 2019; Fan et al., 2021; Paccotacya-Yanque et al., 2022). This way, we prioritized data quantity over in-depth annotation. This approach would ultimately benefit the performance of our classifiers. The final amount of labeled data was determined by the budget allocated by each company for this task.

With the aim of speeding up the annotation process, the collected raw data were first preprocessed with a speaker diarization module (Landini et al., 2022) to separate the clients' and agents' speech. We discarded the segments with speaker overlap as well as segments shorter than 2.5 seconds, which often do not contain enough information to infer the emotional state of the speaker (Tóth et al., 2008).

[1] https://www.mturk.com/

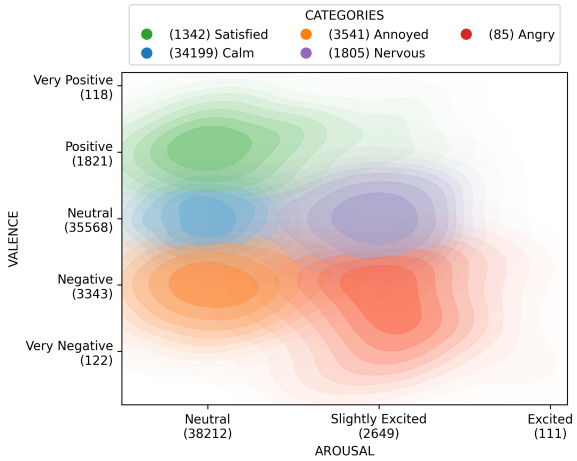


Figure 1: 2D Density Plot of the CC-Support data that shows the most prominent categories in the VA space.

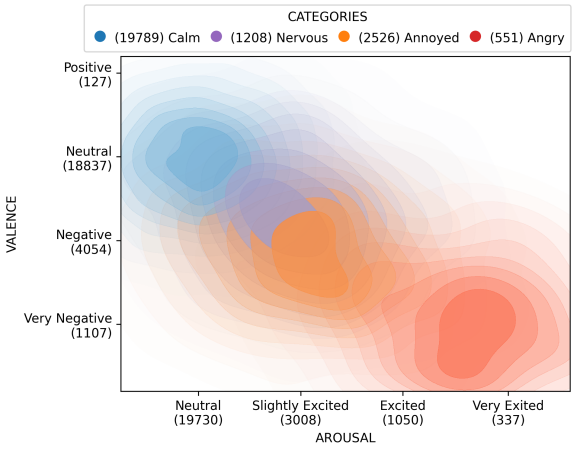


Figure 2: 2D Density Plot of the CC-Debt data that shows the most prominent categories in the VA space.

Finally, we split long turns into several less-than-20-seconds-long segments. This whole process led to an average of 7.3 and 9.1 seconds length segments for CC-Support and CC-Debt, respectively.

Once the preprocessed material from the raw data was prepared, several meetings with the CC annotators were held to establish the labelling criteria. The manual annotation was carried out through an application based on Praat (Boersma, 2001). Regarding the emotion representations, we defined them through the categorical (Ekman et al., 1999; Plutchik, 1980) and the Valence-Arousal dimensional model (Russell, 1980), which was discretized as in (de Velasco et al., 2022). We defined these choices to label each segment:

- Categories: Calm, Nervous, Angry, Annoyed, Surprised, Satisfied.
- Valence: Very Negative, Negative, Neutral, Positive, Very Positive.
- Arousal: Neutral, Slightly Excited, Excited, Very Excited.

2.3 Data preparation and analysis

Once the manual annotation was performed, we computed some data statistics to establish the ground truth labels, defining the following criteria in order to take out the most of our data:

1. If the amount of instances for a particular class is too low, discard them. For example, we removed the “Surprised" label in both use cases due to there were less than 30 labeled segments (combined annotations). The “Satisfied" class was also excluded from the CC-Debt use case for the same reason.

2. Merge two labels if they show a high correlation. For instance, we merged “Annoyed" with “Angry" in both cases.

3. Finally, speech samples where the two annotations differed was further analyzed in a process that involved the experts of the CCs[2]. The vast majority of disagreements were Neutral vs. Emotional labels. In these cases, we selected the Emotional label as the ground truth. Alternatives like discarding these samples, led to overall worse results.

In order to analyze the relation between the categorical emotions, we computed density plots in the arousal-valence plane, as shown in Figures 1 and 2. The numbers in brackets indicate the amount of segments per label (before merging). These maps reveal even more information, such as Annoyed and Angry were very related, in both cases. This phenomena was further noticed in preliminary classification experiments, where there was a noticeably high confusion between both emotions. Therefore, we merged them for our experiments. Similarly, we combined the Positive/Very Positive valences into a single label, the Slightly Excited/Excited arousal

[2]This was only needed for a small proportion of the dataset, because the overall agreement accuracy between the two annotations was 0.82 for CC-Support and 0.90 for CC-Debt, with average Cohen’s kappa coefficients of 0.16 and 0.67, respectively.

Table 1: Amount of data after annotation and post-processing for the CC-Support dataset. Note that some labels have been merged, and others discarded. Complete (Full) and binary (Bin) settings are indicated.

Dim.	Label	Samples	Time (h)	
			Full	Bin
Category	Calm	14817	30.2	30.2
	Annoyed	2723	6.2	
	Nervous	1717	3.2	11.3
	Satisfied	1241	1.9	
Valence	Neutral	15754	31.9	34.7
	Positive	1718	2.8	
	Negative	3026	6.8	6.8
Arousal	Neutral	18025	36.7	
	Excited	2473	4.8	

Table 2: Amount of data after annotation and post-processing for the CC-Debt dataset. Note that some labels have been merged, and others discarded. Complete (Full) and binary (Bin) settings are indicated.

Dim.	Label	Samples	Time (h)	
			Full	Bin
Category	Calm	9476	23.0	23.0
	Annoyed	1717	5.1	7.5
	Nervous	873	2.4	
Valence	Neutral	9039	21.8	21.8
	Negative	2379	6.8	8.7
	Very Neg.	651	1.9	
Arousal	Neutral	10145	22.4	22.4
	Slig. Exc.	2030	5.8	8.1
	Excited	749	2.3	

labels in CC-Support, and the Positive/Neutral valence labels and Excited/Very Excited arousal labels in the CC-Debt use case.

2.4 Final datasets

After the described post-process, we ended up with the amount of data shown in Tables 1 and 2 for CC-Support and CC-Debt, respectively. The information is given per dimension and class. The total amount of valid labeled hours for the CC-Support and CC-Debt use cases reached 41.5 and 30.5 hours respectively, where Calm and Neutral predominated over the emotional classes.

Finally, since one of the main objectives of the SER technology is to detect conflict points during the calls, we also created a more straightforward dataset, where each dimension (categorical, valence, and arousal) is composed of only two classes: Neutral and Emotional. To this end, we kept the majority class as Neutral, whereas the minority classes were merged into the Emotional class. We also indicated in Tables 1 and 2 the resulting amount of hours for this binary setting. This strategy is intended to improve the performance of the models by not only reducing the number of classes but also the imbalance of the data. This decision was jointly taken with the CC experts, as it was determined that different binary classifiers per dimension would provide enough information to assess whether a call should be carefully analysed.

3 Experimental results

3.1 Experimental framework

The constructed SER systems were evaluated on the real CC databases, in addition to the IEMOCAP corpus (Busso et al., 2008), in order to compare their performance in a well-established research database in the community. IEMOCAP consists of five dyadic sessions with two actors (male and female), summing up speech recordings that last nearly 12 hours. Following previous works (Pepino et al., 2021), we only evaluated categorical classification considering four different emotional classes: Anger, Happiness, Neutral, and Sadness.

All the experiments and evaluations were performed with a 5-fold cross-validation technique. For our in-domain databases, we split the recordings into five separate sessions, ensuring balanced (stratified) labels and that the audio samples from the same conversation were not distributed in different folds. Regarding the IEMOCAP corpus, each fold corresponded to a different recording session.

Different kinds of input features were evaluated during our analysis. Traditional features in the SER research community were employed, including eGeMAPS (Eyben et al., 2015), Compare 2016 (Eyben et al., 2015), as well as prosodic features (Parra-Gallego and Orozco-Arroyave, 2022). We also considered SOTA deep features for the SER task. First, we evaluated x-vector embeddings from a ResNet trained for speaker verification (Landini et al., 2022) due to the capability of these models to summarize various paralinguistic factors.

Table 3: F1 results (and CI) for the evaluation on the IEMOCAP dataset. CI of results in **bold** overlap with that of the best resulting model (**underlined**).

Feature	Classifier	IEMOCAP
eGeMAPS	SVM	56.11 ± 3.31
ComPare16		59.35 ± 2.67
Prosody		46.51 ± 2.04
x-vector		59.39 ± 2.83
W2V2		**71.22 ± 2.48**
WavLM		**72.75 ± 2.83**
W2V2	DNN-SP	**71.52 ± 2.54**
	DNN-AttCP	68.07 ± 2.52
WavLM	DNN-SP	**71.98 ± 2.58**
	DNN-AttCP	**73.80** ± **2.35**

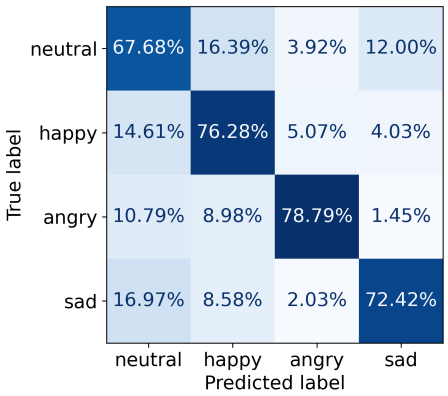


Figure 3: Confusion matrix for the IEMOCAP dataset using a WavLM feature extractor with DNN-AttCP downstream classifier.

Following the current trends, embeddings from self-supervised models such as Wav2Vec2 XLS-R (W2V2) (Babu et al., 2022) and WavLM (Chen et al., 2022) were also analyzed due to their SOTA performance. Finally, the usefulness of content information was also analyzed using linguistic features. To this end, the audio was first transcribed using a medium Whisper model (Radford et al., 2023) fine-tuned with 500 hours of Spanish telephonic speech. The obtained transcriptions were fed to a Spanish BERT model called BETO (Cañete et al., 2020) to compute contextual representations.

Previous features were used to train and test two different machine learning classifiers. The first model consists of support vector machines (SVM) using one-vs-rest classification. The SVM was trained using the radial basis kernel and a balanced class weighting. Moreover, the features were standard normalized using the training set statistics. An average pooling was done for the deep learning models that output temporal sequences to compute the utterance vector representation. Moreover, for the speech self-supervised models, we also considered the hidden layer representations, which are known to contain more discriminative paralinguistic information (wen Yang et al., 2021).

To complete our analysis, DNN downstream classifiers were trained on top of the speech self-supervised models. We followed the approach presented in (Stafylakis et al., 2023; Kakouros et al., 2023), based on pre-trained self-supervised models with a weighted sum of the hidden representations before feeding the fine-tuned downstream network. Two different classifiers were considered, both of them based on embedding computation and softmax classification. The first one performs a linear transformation for dimensionality reduction followed by a simple mean-std statistical pooling (SP), and it is trained using cross-entropy (CE) loss. On the other hand, the second classifier also considers channel dropout and an alternative attentive correlation pooling (AttCP). To compute the attention weights, multiple heads are employed, and the similarities are aggregated prior the corresponding softmax layer via LogSumExp function. Finally, the classifier is trained using the CE loss with label smoothing.

During training, an 80-20 train-development partition was considered for model validation. The ADAM optimizer was used with a learning rate of $3 \cdot 10^{-4}$. Finally, to overcome the imbalanced dataset issue, a down-sampling strategy was followed to reduce samples at each epoch and keep a balanced distribution.

3.2 Results and analysis

We evaluated the different approaches in the IEMOCAP and the CC corpora. For the former, a category-level classification system was built. Similarly, we built one classifier per dimension (i.e., categorical, arousal, and valence) with the CC corpora. Besides, results are reported for both complete and binary label settings. The approaches were evaluated in terms of the macro-averaged F1-score, which accounts for imbalanced datasets. To consider the statistical significance when comparing systems, we performed bootstrapping (Keller et al., 2005; Ferrer and Riera) on the pooled test results to obtain 95% confidence intervals (CI).

Table 4: F1 scores for the evaluation on the CC in-domain data considering the different dimensions. Both the complete and binary version scores are included (separated by /). CI of results in **bold** overlap with that of the best resulting model (**underlined**).

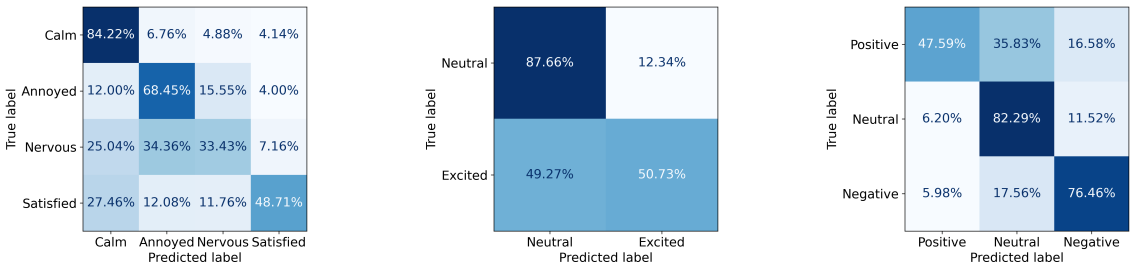
Feature	Classifier	CC-Debt			CC-Support		
		Category	Arousal	Valence	Category	Arousal	Valence
-	Random	26.25 / 45.60	26.40 / 46.05	26.70 / 46.60	19.15 / 47.35	41.60	26.75 / 42.90
	Majority	29.33 / 43.99	29.00 / 43.49	28.55 / 42.82	20.98 / 41.96	46.79	28.97 / 46.01
eGeMAPS	SVM	56.07 / 76.66	54.77 / 74.73	58.57 / 77.58	43.68 / 75.21	60.95	50.96 / 67.92
ComPare16		58.53 / 79.15	**56.18 / 77.28**	59.71 / 79.24	46.43 / 76.59	62.23	55.62 / 71.40
Prosody		49.82 / 72.22	49.54 / 70.90	51.07 / 72.10	37.58 / 68.86	57.71	46.04 / 63.89
x-vector		57.68 / 78.51	55.21 / 75.42	57.79 / 77.12	48.20 / 79.31	**64.41**	55.82 / 70.93
BETO		54.68 / 74.49	50.14 / 72.29	53.47 / 76.52	51.74 / 74.83	61.88	60.94 / **74.53**
W2V2		**64.35 / 82.63**	**62.05 / 80.32**	**66.18 / 82.41**	**56.04 / 81.78**	**66.14**	**64.43 / 77.08**
WavLM		**63.69 / 81.65**	**60.67 / 79.83**	**64.83 / 81.30**	**55.00 / 81.67**	**66.25**	**62.87 / 76.07**
W2V2	DNN-SP	**63.94 / 82.71**	**61.11 / 80.08**	**64.63 / 81.57**	**56.62 / 81.04**	**66.44**	**64.35 / 74.94**
	DNN-AttCP	**61.46 / 81.09**	**60.11 / 78.00**	**61.48 / 80.33**	**54.39** / 79.57	**64.60**	**62.47** / 73.95
WavLM	DNN-SP	**64.53 / 82.27**	**60.53 / 79.51**	**63.07 / 81.26**	**56.36 / 81.23**	**66.63**	**64.12 / 75.05**
	DNN-AttCP	**62.61 / 81.14**	**59.79 / 79.00**	**62.45 / 80.22**	**54.81 / 80.41**	**64.89**	**63.20 / 74.70**

We first considered the IEMOCAP results in Table 3 to evaluate the different systems in an established benchmark. As it can be clearly noted, in this case the best performance is obtained using deep features from self-supervised acoustic models. Nevertheless, there is no statistical difference between using W2V2 and WavLM as feature extractors when comparing SVM and DNN classifiers. Moreover, the results using the DNN downstream models are comparable with those reported in (Kakouros et al., 2023), which are SOTA metrics in the speech-only benchmark. Thus, the analysis of these results pointed out that the main improvements come from these self-supervised models as feature extractors. At the same time, a simple SVM classifier is robust enough to exploit the paralinguistic information of the deep embeddings to perform SER classification.

Figure 3 shows the confusion matrix for the IEMOCAP dataset using the WavLM feature extractor with DNN-AttCP downstream classifier. It can be observed that the per-class accuracy ranges between 70%-80%, except for the neutral class, which shows the lowest per-class accuracy (68%). Moreover, a high percentage of misclassifications are observed between the neutral and the remaining emotions, which could be expected in this task when the system does not clearly detect the emotion in the speech signal. Indeed, the results are consistent with those obtained by state-of-the-art recent SER studies (Kakouros et al., 2023; Ulgen et al., 2024; Shome and Etemad, 2024).

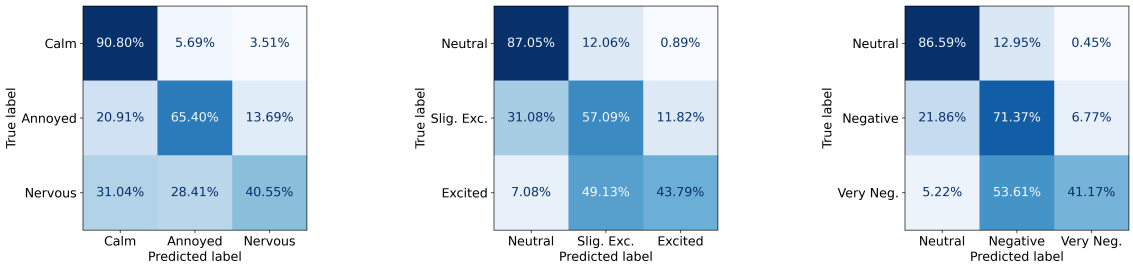
Regarding the in-domain CC datasets, similar tendencies are found. Table 4 shows the experimental results for the CC-Debt and CC-Support datasets. We also included results obtained when using two baseline classifiers: a random classifier (results averaged over 50 trials) and a majority voting classifier. As observed, the best performance is obtained using W2V2 and WavLM features regardless of the classifier. These results confirm the well-known capabilities of self-supervised models for this task. As expected, the label merging process improved different use cases and dimensions, with F1 scores close to suitable values for practical applications (Płaza et al., 2022; Deschamps-Berger et al., 2023). For the case of CC-Debt, the gains obtained for the complete vs. binary level configurations are similar across the different dimensions, showing possible correlations among them.

On the contrary, the disparity is higher in CC-Support. The observed improvements in categorical classification are probably due to the label reduction process. Nevertheless, the F1 metrics are lower in the case of valence and, especially, arousal, where there were two classes from the initial version. The discrepancy with respect to the CC-Debt is that the former is a less-emotional domain, and the data for the minority classes were scarce. Therefore, accurately detecting excitement and negative emotions was even more challenging. Still, the results are competitive, considering the task's complexity (Deschamps-Berger et al., 2023). Other features, such as ComPare16 or x-vectors,



(a) Category (b) Arousal (c) Valence

Figure 4: Confusion matrices for the CC-Support dataset using a W2V2 feature extractor with SVM classifier. These matrices are obtained by dimension (Category, Arousal, Valence) in the complete classes setting.



(a) Category (b) Arousal (c) Valence

Figure 5: Confusion matrices for the CC-Debt dataset using a W2V2 feature extractor with SVM classifier. These matrices are obtained by dimension (Category, Arousal, Valence) in the complete classes setting.

show strong results for particular cases. The linguistic features exhibit higher accuracy in valence prediction within the CC-Support domain, likely attributable to the relative ease of transcription due to the standardized vocabulary prevalent in these conversations.

To further analyze these results, Figures 4 and 5 show the confusion matrices obtained for the CC-Support and CC-Debt datasets, respectively, when evaluating the W2V2 feature extractor with SVM classifier. We only show the results for the complete classes setting as it yields a better understanding of the main errors produced by the systems. For the CC-Support, the detection of emotion categories such as Nervous and Satisfied has a low detection rate due to the few amount of hours, which justifies using a simple binary detection between neutral and emotional classes, with most of the cases representing Annoyed or Nervous users. Similarly, Positive valences are mainly confused with Neutral, while detecting Negative vs others can bring better discriminative results that help detect these altered states. For the CC-Debt, similar behavior is observed for the emotional category detection. Regarding arousal and valence dimensions, extreme classes are mainly confused with the adjacent intermediate level (e.g., Excited with Slightly Excited arousal and Very Negative with Negative valences). Thus, in this scenario, with few labeled data for these classes, it is justified to simplify the detection problem and group the non-neutral levels in a single class while keeping the usefulness of the deployed systems.

In conclusion, using self-supervised features from large speech models with classical machine learning classifiers such as SVMs can obtain promising results for a practical application of SER when considering a simplified scenario focused on detecting Neutral and Emotional classes. Moreover, in more complex scenarios with class variety, the results are still competitive regarding the current state-of-the-art in this area (Kakouros et al., 2023; Deschamps-Berger et al., 2023). It is also important to remark that using general SSL feature extractors trained on a large variety of speech data avoids the need for transfer learning from pre-trained SER models (e.g., IEMOCAP), especially when there is a considerable domain shift between the source and target scenarios (different languages, acoustic channel, acted vs. real emotions).

3.3 Subjective analysis

Finally, we conducted a subjective evaluation to further analyze the usability of the SER models in a practical scenario for customer service and speech analytics. To this end, a new test set of 15 recorded conversations from the CC-Debt domain is considered, including both neutral and emotional states. Two different evaluators are involved in the task. The first one listens to the conversations and takes notes about their emotional content and evolution during the call, both for the agent and the client. The second one only analyzes the information provided by the automatic pipeline, which includes segmentation, diarization, and emotion recognition (including both full and binary models). Then, the evaluator described the conversation's emotional evolution using only this information. Finally, the evaluators compared their analysis.

After finalizing this procedure, five conversations were categorized as generally neutral, and the remaining ones as emotional. The analysis of both evaluators matched in 100% of the conversations, regarding general aspects as the evolution of emotional state, considering both agent and client during the conversation. Despite minor errors not only related to the emotional models but also other modules in the pipeline (e.g., speaker segmentation by the diarization step), the automatic analysis allowed us to obtain an overview of the call and evolution of emotional states. Interestingly, the second evaluator remarked on the usefulness of multi-class models for arousal and valence to evaluate the temporal evolution and accurately identify segments with strong negative emotions (excited arousal and very negative valence). On the other hand, using binary models for categorical prediction was preferred to identify the negative state. To summarize the outcomes from this evaluation, the involved company identified practical use cases where the emotional models, along with other speech-related technologies (such as automatic speech recognition and content classification) have potential applicability, including: (1) call identification with very negative emotions from the client (especially exploiting extreme categories for arousal and valence), (2) analysis of emotional evolution on these calls, and (3) evaluation of agents performance, where mid-level emotional states are important to analyze the conversations' emotional evolution.

4 Conclusions

This paper presented a practical technological transfer of speech emotion recognition systems to the CC speech analytics sector. This work results from collaboration between research teams and two CC-related companies to address specific target scenarios. In this study, we completed all the necessary stages to ensure that the systems are production-ready from their facilities: data acquisition, preprocessing and annotation, analysis and design of the experimental framework, training, and evaluation of the different approaches. The CC-expert annotation process ensures the quality of the data while meeting all the privacy concerns, which usually causes several issues in developing these systems. Moreover, we considered two application scenarios to evaluate the original annotated data and a transformed version (2 classes) focused on detecting emotional states. The experimental results indicated that the proposed approaches are competitive for the two scenarios, as well as in a well-established benchmark in the research community. This work represents a successful technological transfer to the industry, where the companies have deployed the solutions to evaluate it in their commercial cases. In future work, we will study these models' use along with active learning techniques to help annotate additional emotional data.

Limitations

The main limitation of this work is the generalization and application of the development of SER systems in out-of-domain conditions. The models have been trained on a limited amount of labeled speech in specific conditions of language, acoustic channel, application domain, and targeted emotions. Thus, using these models under different conditions will result in a performance drop and non-sense results. Therefore, these systems should only be considered under similar conditions.

Another limitation is the dependence of the SER module on previous speech-processing steps in real applications, including speech segmentation and speaker diarization. Thus, errors in the previous steps of the pipeline will ultimately affect the predictions obtained by the SER system.

Ethics Statement

The EU AI Act considers systems that predict human emotions from their biometric data. There

is a concern about them due to their potential biases and lack of generalization, as well as their potential to limit rights and freedom for human beings. Thus, these systems are generally considered high-risk and strictly forbidden in domains where specific person profiles are targeted, such as work and education environments. Using these systems in sensitive domains should only be allowed with a healthcare objective or to ensure people's security.

In the context of analyzing call-center conversations, the use of a SER system may be classified as a non-high-risk application, as long as it ensures the protection of individuals' health, security, and human rights. Additionally, it is critical to implement measures that prevent potential biases in the AI models and ensure that these systems do not significantly influence decision-making processes, which should always be reviewed by human experts. Furthermore, this technology must not be used to profile clients within this domain, and the results should always be anonymized to protect individual privacy. Moreover, with regard to transparency obligations, clients should always be informed that their conversations are being recorded and analyzed using these AI systems, and they should be given the right to object to these operations. Finally, the deployment of SER systems in CC scenarios should always be carried out under the supervision of ethics experts to ensure compliance with rules and directives outlined in EU regulations.

Acknowledgements

This work was partially funded by the Basque Business Development Agency, SPRI, under the EMPHASIS project, grant agreement ZE-2021/00039.

References

Arun Babu et al. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Proc. InterSpeech 2022*, pages 2278–2282.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345.

Milana Bojanić, Vlado Delić, and Alexey Karpov. 2020. Call redistribution for a call center based on speech emotion recognition. *Applied Sciences*, 10(13):4653.

Carlos Busso et al. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Sanyuan Chen et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Mikel de Velasco, Raquel Justo, and María Inés Torres. 2022. Automatic identification of emotional information in spanish tv debates and human–machine interactions. *Applied Sciences*, 12(4):1902.

Théo Deschamps-Berger, Lori Lamel, and Laurence Devillers. 2021. End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings. In *Proc. 2021 ACII*, pages 1–8.

Theo Deschamps-Berger, Lori Lamel, and Laurence Devillers. 2023. Exploring attention mechanisms for multimodal emotion recognition in an emergency call center corpus. In *Proc. ICASSP 2023*.

Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Florian Eyben et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Weiquan Fan, Xiangmin Xu, Xiaofen Xing, Weidong Chen, and Dongyan Huang. 2021. LSSED: A large-scale dataset and benchmark for speech emotion recognition. In *Proc. ICASSP 2021*, pages 641–645.

Yajing Feng and Laurence Devillers. 2023. End-to-end continuous speech emotion recognition in real-life customer service call center conversations. In *Proc. 2023 ACII Workshop and Demos*.

Luciana Ferrer and Pablo Riera. Confidence intervals for evaluation in machine learning.

C Hema and Fausto Pedro Garcia Marquez. 2023. Emotional speech recognition using CNN and deep learning techniques. *Applied Acoustics*, 211:109492.

Christian Hildebrand, Fotis Efthymiou, Francesc Busquet, William H Hampton, Donna L Hoffman, and Thomas P Novak. 2020. Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications. *Journal of Business Research*, 121:364–374.

Jon Irastorza and M. Ines Torres. 2016. Analyzing the expression of annoyance during phone calls to complaint services. In *Proc. IEEE CogInfoCom*, pages 103–106.

Jon Irastorza and M. Ines Torres. 2019. Tracking the expression of annoyance in call centers. *Cognitive Infocommunications, Theory and Applications*, pages 131–151.

Sofoklis Kakouros, Themos Stafylakis, Ladislav Mošner, and Lukáš Burget. 2023. Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing. In *Proc. ICASSP 2023*.

Mikaela Keller, Samy Bengio, and Siew Wong. 2005. Benchmarking non-parametric statistical tests. *Advances in neural information processing systems*, 18.

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71:101254.

Manon Macary, Marie Tahon, Yannick Estève, and Daniel Luzzati. 2023. Acoustic and linguistic representations for speech continuous emotion recognition in call center conversations. *arXiv preprint arXiv:2310.04481*.

Abdelrahman Mohamed et al. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Rosa YG Paccotacya-Yanque, Candy A Huanca-Anquise, Judith Escalante-Calcina, Wilber R Ramos-Lovón, and Álvaro E Cuno-Parari. 2022. A speech corpus of quechua collao for automatic dimensional emotion recognition. *Scientific Data*, 9(1):778.

E Parada-Cabaleiro, G Costantini, A Batliner, A Baird, Bw Schuller, et al. 2018. Categorical vs dimensional perception of italian emotional speech. In *Proc. InterSpeech 2018*, volume 2018, pages 3638–3642.

Luis Felipe Parra-Gallego and Juan Rafael Orozco-Arroyave. 2022. Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments. *Digital Signal Processing*, 120:103286.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *Proc. InterSpeech 2021*, pages 3400–3404.

Paula Andrea Pérez-Toro, Juan Camilo Vásquez-Correa, Tobias Bocklet, Elmar Nöth, and Juan Rafael Orozco-Arroyave. 2021. User state modeling based on the arousal-valence plane: Applications in customer satisfaction and health-care. *IEEE Transactions on Affective Computing*, 14(2):1533–1546.

Mirosław Płaza et al. 2022. Emotion recognition method for call/contact centre systems. *Applied Sciences*, 12(21):10951.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, pages 28492–28518.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Debaditya Shome and Ali Etemad. 2024. Speech emotion recognition with distilled prosodic and linguistic affect representations. In *Proc. ICASSP 2024*, pages 11976–11980.

Themos Stafylakis, Ladislav Mošner, Sofoklis Kakouros, Oldřich Plchot, Lukáš Burget, and Jan Ćernockỳ. 2023. Extracting speaker and emotion information from self-supervised speech models via channel-wise correlations. In *Proc. IEEE SLT 2022*, pages 1136–1143.

Szabolcs Levente Tóth, David Sztahó, and Klára Vicsi. 2008. Speech emotion perception by human and machine. In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction: COST Action 2102 International Conference*, pages 213–224.

Ismail Rasim Ulgen, Zongyang Du, Carlos Busso, and Berrak Sisman. 2024. Revealing emotional clusters in speaker embeddings: A contrastive learning strategy for speech emotion recognition. In *Proc. ICASSP 2024*, pages 12081–12085.

Mikel deVelasco Vázquez, Raquel Justo, Asier López Zorrilla, and Maria Inés Torres. 2019. Can spontaneous emotions be detected from speech on tv political debates? In *10th IEEE International Conference on Cognitive Infocommunications*, page 289.

Shu wen Yang et al. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. InterSpeech 2021*, pages 1194–1198.

Fangfang Zhu-Zhou, Roberto Gil-Pita, Joaquín García-Gómez, and Manuel Rosa-Zurera. 2022. Robust multi-scenario speech-based emotion recognition system. *Sensors*, 22(6):2343.