Performance Gaps in Acted and Naturalistic Speech: Insights from Speech Emotion Recognition Strategies on Customer Service Calls

Lily Kawaoto, Hita Gupta, Ning Yu, and Daniel Dakota
Leidos Holdings, Inc.
{lily.kawaoto,hita.gupta
ning.yu,daniel.d.dakota}@leidos.com

Abstract

Current research in speech emotion recognition (SER) often uses speech data produced by actors which does not always best represent naturalistic speech. This can lead to challenges when applying models trained on such data sources to real-world data. We investigate the application of SER models developed on acted data and more naturalistic podcasts to service call data, with a particular focus on anger detection. Our results indicate that while there is noticeable performance degradation of models trained on acted data to the naturalistic data, weighted multimodal models developed on existing SER datasets–both acted and natural–show promise, but are limited in ability to recognize emotions that do not discernibly cluster.

1 Introduction

Speech emotion recognition (SER) aims to enhance interactive speech systems to pick up on emotional cues in a user's voice. The applications range in use from customer care, robotics, and education applications among others. By being more sensitive to a user's emotional state, a better user experience can be provided. For example, detection of increasing anger in an automated voice support system can prompt an immediate transfer to a human agent, helping to quickly de-escalate the situation and potentially improving customer feedback.

Much SER research focuses on improving emotion in a multi-class setting rather than focusing on one emotion (Han et al., 2014; Fayek et al., 2017; Pepino et al., 2021). More recently, efforts to generalize SER from multilingual and multi-corpus angles have increased as well (Radford et al., 2022; Ma et al., 2024; An et al., 2024). However, the most commonly used open-source SER dataset, IEMOCAP (Busso et al., 2008), consists of improvised or scripted speech from actors. This leaves questions on how generalizable the findings are as models derived from IEMOCAP may not always match emotional class distributions on real world data, such as anger (Erden and Arslan, 2011; Pappas et al., 2015), which is often more relevant from a customer service perspective.

We focus on investigating the application of SER strategies with noted positive performance on acted emotion dataset to more naturalistic data. Specifically, we first replicate Chen and Rudnicky (2021) who mainly used the IEMOCAP dataset to compare a traditional fine-tuning approach to two continued pre-training techniques. We then extended this study with additional experiments focusing on the use of IEMOCAP in cross-corpus experiments using the MSP dataset (which consists of podcast recordings; Lotfian and Busso, 2019) as well as a small set of real world data from an IT customer service center. Different factors including model architecture, single- vs multi-modality, and cross-corpus training are introduced in our experiments to thoroughly investigate how effectively techniques can be applied to improve performance on more naturalistic speech.

2 Related Work

2.1 Transformer Models in SER

Recent research in SER has shifted focus to using transformer-based models (Chen and Rudnicky, 2021; Wagner et al., 2022), including SOTA models such as Whisper (Radford et al., 2022), WavLM (Chen et al., 2022), and Emotion2Vec (Ma et al., 2023). In our pilot study however we use the Wav2Vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) models as a starting point to reproduce prior fine-tuning approaches and explore their generalizability to naturalistic data.

Proceedings of Recent Advances in Natural Language Processing,pages 521–530
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-061

2.2 SER Techniques

Domain Adaptive Pre-training (DAPT) is continued pre-training with the same self-supervised objective functions as the base pre-trained model. Such an approach has been shown to help adapt a general LLM to the target domain with unlabeled target data. Gururangan et al. (2020), for example, compared performance of a base transformer model across various domains to its domain-adapted (DAPT) version as well as its irrelevant domain-adapted (¬DAPT) version. The DAPT models consistently outperformed the base model on all domains whereas the ¬DAPT models generally showed worse performance than the base model, indicating that exposure to relevant domain data positively impacts results on the end task. However, the challenge with applying DAPT to SER is that DAPT only learns general features in audio; it does not explicitly capture emotional context (Gururangan et al., 2020; Liu et al., 2021).

Pseudo-label Task Adaptive Pre-training (PTAPT) is an extension of task-adaptive pre-training (TAPT). TAPT uses continued pre-training with unlabeled data that is smaller than what is used for DAPT but is more task-relevant and similar to the downstream tasks. Chen and Rudnicky (2021) extend this to PTAPT which has outperformed both DAPT and fine-tuning alone on IEMOCAP. PTAPT uses k-means clustering on utterance-level labels to generate "pseudo-labels" for frame-level emotion representation, addressing label granularity by predicting the pseudo-labels, instead of the masked audio frames, during pre-training. Importantly, the same data samples are used in the continued pre-training and fine-tuning processes, meaning emotion labeled data is used both with labels known in fine-tuning and unknown in DAPT or TAPT, which distinguishes it from previous domain and adaptation experiments.

Multimodal Experiments from multiple source representations, such as leveraging both the audio and corresponding transcripts, have been shown to yield superior SER results than using audio alone. Concatenation is a common approach for dealing with multimodal SER features. Feng et al. (2020) implement a concatenation technique but use the hidden state of the decoder in an ASR model to replace word embeddings as inputs to an SER model, effectively utilizing only speech features yet achieving performance similar to that of multimodal speech-and-text models. Hazarika et al. (2022) proposed a multimodal network to account for multi-speaker utterances. Each speaker is represented by a concatenation of text transcript, spatiotemporal (facial expression), and audio features. A self-influence module then represents each speaker's emotional dependency by accounting for their previous utterance states. All the speakers' features are processed temporally by a dynamic global influence module to maintain a global representation of the conversation. Finally, a multi-hop memory module refines the features into a context-aware representation for emotion prediction.

Probol and Mieskes (2023) take a different feature combination approach. After creating nine audio neural network models with varying internal specifications and training each on speech data, two combinations were explored: combining one audio model with either another audio model or the text model, each with a weight of 50%, and likewise combining three models, each with a weight of 33%.

3 Methodology

3.1 Data

IEMOCAP (Busso et al., 2008) consists of American English speech from either scripted or improvised recordings by ten actors. It was recorded in five dyadic sessions (i.e., five interactions with two actors each). Due to label sparsity of some emotional categories, we only use the "neutral", "happy", "sad", and "anger" emotion labels, combining "excited" with "happy" following Chen and Rudnicky (2021). The number of samples and distribution percentages per emotion class for this modified version of IEMOCAP is shown in the first row of Table 1. Each audio file has an accompanying text transcription.

MSP Podcast (Lotfian and Busso, 2019) contains segments in various English dialects from online podcasts covering a variety of topics (e.g., political debates and movie reviews), thus is more representative of realistic speech. We use the accompanying podcast transcripts and the same four emotional categories as IEMOCAP. An important difference however is MSP's severely imbalanced class distribution compared to IEMOCAP. As seen in Table 1, the majority of samples are labeled "neu-

Dataset	Split	anger	happy	neutral	sad	total
IEMOCAP	PT/FT	893 (20%)	1337 (30%)	867 (19%)	1373 (31%)	4470 (100%)
	Test	210 (20%)	299 (28%)	335 (32%)	217 (20%)	1061 (100%)
MSP	PT	3178 (9%)	10990 (30%)	20563 (55%)	2480 (7%)	37211 (100%)
	FT	364 (8%)	1359 (30%)	2437 (55%)	310 (7%)	4470 (100%)
	Test	65 (6%)	350 (33%)	590 (56%)	56 (5%)	1061 (100%)
MSP*	PT/FT	893 (20%)	1337 (30%)	867 (19%)	1373 (31%)	4470 (100%)
	Test	210 (20%)	299 (28%)	335 (32%)	217 (20%)	1061 (100%)
Service Desk	Test	7 (2%)	8 (3%)	283 (95%)	0 (0%)	298 (100%)

Table 1: Number of samples over all emotion classes for datasets used in this study.

tral" (55% in the pre-train / fine-tune sets), followed by "happy" (30%) and a sharp decrease for anger (9%) and sad (7%). For this reason, we also create a randomly selected balanced subset, **MSP***, with a similar size and class distribution to IEMOCAP.

Service Desk Calls consist of six approved conversations of proprietary recorded service desk calls between service desk agents and customers requesting IT help in American English.[1] An authorized service desk team was permitted access to the original audio calls, which then underwent manual removal of all PII and generated transcripts using Google Cloud's Speech-to-Text API. After encrypted distribution of these modified files to the authors, the data was stored in a private AWS S3 bucket with strict access policies. Transcripts of the calls did not contain timestamps; for compatibility with our multimodal experiments we used Aeneas[2], a forced aligner, to create utterances from the modified audio and transcripts, which were then manually annotated for the four emotion labels by one annotator and then reviewed by a second. While we recognize the number of speakers is limited for this pilot, this is only used as a test set. Annotating each call at the utterance level yielded 298 utterances: 7 "anger", 8 "happy", 283 "neutral", with an average utterance length of 5.6 seconds.

3.2 Models

We compare the performance of two models, Wav2Vec 2.0 and HuBERT (Wagner et al., 2022; Chang et al., 2021).[3] Wav2Vec 2.0 learns to recognize discrete units in waveforms through self-supervised training via a codebook of speech units coupled with a quantization process, while HuBERT differs by using unsupervised clustering of MFCC features to learn speech representations in the training step. The Fairseq library[4] was used to perform the HuBERT DAPT experiments using the same learning rate and number of training steps as Wav2Vec 2.0 DAPT, since a pre-training model for HuBERT was not available on HuggingFace at the time of our experiments. Chen and Rudnicky (2021) use Wav2Vec 1.0 for the clustering stage in their PTAPT pre-training; we use the version provided by the Fairseq library. For PTAPT, we apply the same clustering method for both models.

We used distributed data processing (DDP) and gradient accumulation in all pre-training and fine-tuning stages to achieve an effective batch size of 64 with a learning rate of 1e-4. Fine-tuning ran for 15 epochs, the pseudo-label clustering ran for 300k training steps, and the continued pre-training stage for both DAPT and PTAPT ran for 120k training steps. We take the average of five-fold evaluations to obtain a generalized score for each experiment. Approximate times for each fold (time ranges are due to dataset size variations, with IEMOCAP and MSP* completing faster than MSP): 1-2 hours for vanilla fine-tuning (on a g4dn.12xlarge (4 GPUs)), 3-5 hours for DAPT, and 4-6 hours for PTAPT (both using a p3.16xlarge (8 GPUs)).

3.3 Experimental Design

Single Corpus experiments examine if PTAPT outperforms DAPT and fine-tuning alone with naturalistic data, using the Wav2Vec 2.0 and HuBERT modelson audio features only. To establish baselines, we replicate the same single-corpus experiments described in Chen and Rudnicky (2021) separately on IEMOCAP and MSP, then compare the PTAPT, DAPT, and fine-tuning only (FT-only) results. To make fairer performance comparisons

[1]This data received blanket approval for use in this task.

[2]https://github.com/readbeyond/aeneas

[3]We use the HuggingFace API to obtain both base models: facebook/wav2vec2-base, facebook/hubert-base-ls960

[4]https://github.com/facebookresearch/fairseq

Pretrain-Finetune-Test sets	Wav2Vec 2.0			HuBERT		
	FT-only	DAPT	PTAPT	FT-only	DAPT	PTAPT
IEMO-IEMO-IEMO	74.47	76.91	78.15	76.06	77.99	77.23
MSP-MSP-MSP	38.69	37.26	41.51	39.74	41.85	40.82
IEMO-MSP-MSP	38.69	40.69	40.57	39.74	41.48	41.68
IEMO-IEMO-MSP	40.88	39.67	41.21	38.04	39.08	40.35
MSP*-MSP*-MSP*	47.42	45.23	45.61	46.36	45.4	45.46
IEMO-MSP*-MSP*	47.42	45.65	44.69	46.36	44.18	45.69
IEMO-IEMO-MSP*	38.09	41.19	39.98	40.29	40.59	39.94

Table 2: Unweighted Accuracy results from Wav2Vec 2.0 and HuBERT experiments, on audio-only modality. Red is the lowest score for each row at the Model level, yellow the medium score, and green the highest. Row 1 replicates prior literature on IEMOCAP. Rows 2-7 evaluate on MSP or MSP*.

Pretrain-Finetune-Test sets	FT-only	DAPT	PTAPT	PTAPT improv.
IEMO-IEMO-IEMO	71.72	76.25	78.6	9.59%
MSP-MSP-MSP	58.15	50.61	55.04	-5.34%
IEMO-MSP-MSP	58.15	61.64	55.51	-4.54%
IEMO-IEMO-MSP	45.81	56.93	46.75	2.05%
MSP*-MSP*-MSP*	49.2	50.33	47.79	-2.87%
IEMO-MSP*-MSP*	49.2	47.79	45.05	-2.87%
IEMO-IEMO-MSP*	43.26	43.36	42.97	-0.67%

Table 3: Unweighted Accuracy results on multimodal (audio+text) Wav2Vec 2.0 experiments. Red is the lowest score for each row, yellow the medium score, and green the highest. PTAPT percentage of improvement is calculated against FT-only.

of naturalistic speech to acted speech, we also run an experiment on a downsampled version of MSP (MSP*), which mirrors the sample size and emotion class distribution of IEMOCAP. For all single-corpus experiments, the same data samples are used for pre-training and fine-tuning, and the test data is from the same corpus.

Cross-Corpus experiments examine the value in using models fine-tuned with acted speech data for a naturalistic test set. Because publicly available SER data is very limited, it would be valuable to know whether existing SER data (regardless of whether it is acted or not) can be leveraged for applications to naturalistic speech. Combinations of the model, method, and modality variables are the same as our Single Corpus experiments.

Multimodal experiments extend Chen and Rudnicky (2021)'s study but with a multimodal flavor in which we take the speech model with the better overall performance (i.e., Audio + Text level).

We adopt a cross-representational model by Makiuchi et al. (2021) for our multimodal approach, where probabilities obtained from the better SER model (p_s) and Text-based Emotion Recognition (TER) model (p_t) are respectively multiplied by weights $w_s = 0.75$ and $w_t = 0.6$ respectively.[5] We use the CLS token of `bert-base-case` on the utterance level of the text transcriptions provided with the IEMOCAP and MSP datasets as our text-based features.

Service Desk Calls enable the assessment of these SER solutions on actual customer data to obtain a further accurate picture on their performance on real-life downstream tasks.

Evaluation consists of performing five runs for each experiment and reporting unweighted accuracy (UA)[6] and macro-F1 to better evaluate the unbalanced nature often found in real world scenarios. In all cases, except when replicating Chen and Rudnicky (2021)'s IEMOCAP experiment, we test our models using MSP or MSP* due to the corpus being more representative of naturalistic speech.

4 Results & Discussion

4.1 Single corpus

Table 2 summarizes unweighted accuracy results using the Wav2Vec 2.0 and HuBERT models with single corpus experiments (represented in rows 1, 2, and 5). Our replicated experiment with Wav2Vec 2.0 on IEMOCAP aligned with Chen and Rudnicky (2021): PTAPT outperforms both FT-only and DAPT on IEMOCAP and MSP, while DAPT only showed better performance over fine-tuning, suggesting continued pre-training without taking emotion into consideration may degrade performance.

Results using HuBERT however do not align, as DAPT outperformed PTAPT on both IEMOCAP and MSP. The increased performance seen with PTAPT in Wav2Vec 2.0 may not generalize to other models using only in-corpus data; Wav2Vec 2.0's use of waveforms as targets during self-supervised training may have enhanced PTAPT's psuedo-label prediction step, while HuBERT's use of MFCC clusters may not have.

Both MSP and MSP* show dramatically lower performance drops than IEMOCAP for both

[5]Weights were manually determined after extensive experimentation to determine the combination that yielded the best overall performance across methods. They are the same across all multimodal experiments and since they determine the degree of contribution of each data modality to the final combined probability, are independent of one another and do not need to sum up to one.

[6]This is to align with Chen and Rudnicky (2021) who reported UA in their experiments.

Pretrain-Finetune-Test sets	Emotion	Audio			Audio + Text		
		FT-only	DAPT	PTAPT	FT-only	DAPT	PTAPT
IEMO-IEMO-IEMO	anger	78.09	82.73	**83.86**	74.95	80.16	***85.98***
	happy	72.63	74.38	**76.37**	71.14	***77.8***	77.46
	neutral	65.39	70.85	**71.47**	66.09	67.38	***73.07***
	sad	75.01	**78.68**	78.13	75.7	80.85	***81.51***
MSP-MSP-MSP	anger	26.16	21.85	**30.56**	34.15	23.36	***34.97***
	happy	44.74	50.75	**50.95**	***55.56***	51.73	54.57
	neutral	***67.64***	57.79	67.26	**67.5**	60.85	62.5
	sad	**14.18**	13.68	12.79	***17.39***	15.25	6.19
IEMO-MSP-MSP	anger	26.16	27.05	***41.29***	22.1	23.33	**24.44**
	happy	44.74	51.33	***52.56***	47.58	50.97	**52.46**
	neutral	**67.64**	67.30	47.35	56.19	***68.56***	50.78
	sad	14.18	13.25	**18.34**	18.69	5.26	***20.83***
IEMO-IEMO-MSP	anger	19.40	21.18	***47.75***	**37.19**	30.77	32.75
	happy	47.91	**49.87**	45.69	***56.24***	51.35	53.11
	neutral	**56.56**	55.82	41.37	69.19	***71.84***	64.64
	sad	13.02	15.62	**16.55**	***16.67***	11.49	12.96
MSP*-MSP*-MSP*	anger	**53.92**	52.67	51.07	55.04	55.98	***57.0***
	happy	49.94	**51.41**	50.47	52.78	***55.56***	53.47
	neutral	44.27	46.36	**46.79**	45.53	***49.08***	45.24
	sad	**40.06**	28.67	36.24	***44.94***	38.54	28.66
IEMO-MSP*-MSP*	anger	**53.92**	51.47	51.32	***55.86***	55.42	47.62
	happy	49.92	47.90	**50.07**	***53.05***	39.17	51.13
	neutral	44.27	**46.87**	45.90	46.72	***51.26***	47.13
	sad	**40.06**	37.88	33.30	***44.84***	41.98	23.13
IEMO-IEMO-MSP*	anger	41.83	45.57	**45.75**	47.81	46.22	***48.09***
	happy	43.68	**46.05**	44.10	***50.51***	49.24	46.93
	neutral	42.65	46.36	**47.40**	40.07	44.78	***47.92***
	sad	14.69	**19.37**	9.41	***26.11***	22.59	8.03

Table 4: Class-level performance of macro-F1 scores from experiments using Wav2Vec 2.0, on audio-only and audio+text modalities. In the multimodal experiments, the SER model is pre-trained and fine-tuned on the respective datasets in the first column; the TER model (BERT) is fine-tuned with the union of the two. Highest scores in each row for the two modalities are bolded. Highest scores for anger for each modality is underlined. Higher score per emotion across audio-only and audio+text modalities is italicized.

Wav2Vec 2.0 and HuBERT. While data imbalance certainly plays a role for MSP, MSP* (the balanced subset) tends to support the idea that emotional characteristics in acted speech are intrinsically more exaggerated relative to naturalistic speech (Schuller et al., 2010), and thus easier to recognize. All three continued pre-training methods show improvement from MSP to MSP*, although performance gaps on MSP* compared to IEMOCAP still exist.

4.2 Cross Corpus

Table 2 shows cross corpus experiments (represented in rows 3, 4, 6, and 7). Since the test data in cross corpus experiments differ from either the pre-training or fine-tuning data, results can further reveal the robustness of the continued pre-training methods and suggest the required amount of in-corpus data needed for downstream tasks using naturalistic speech.

Testing on MSP (rows 3-4) shows PTAPT typically yielding higher UAs than FT-only and DAPT, regardless of the fine-tuning dataset and the model. This suggests that PTAPT may be a promising approach if the downstream task involves evaluating on overall performance of naturalistic speech with a class imbalance. Further, the marginal differences between fine-tuning with acted versus natural speech is reassuring when only a small quantity of naturalistic data is available and fine-tuning on in-source data is not viable.

However, when testing on MSP* (rows 6-7), FT-only and DAPT appear to be more robust than PTAPT in both Wav2Vec 2.0 and HuBERT experiments. Similar trends were seen in the single-corpus MSP* experiments, where FT-only outperformed DAPT and PTAPT methods. This may suggest that PTAPT is not as effective on smaller datasets, particularly on naturalistic speech whose emotional features are less emphasized than acted

Pretrain-Finetune-Test sets	Audio			Audio + Text		
	FT-only	DAPT	PTAPT	FT-only	DAPT	PTAPT
IEMO-IEMO-IEMO	78.09	82.73	83.86	74.95	80.16	85.98
MSP-MSP-MSP	26.16	21.85	30.56	34.15	23.36	34.97
IEMO-MSP-MSP	26.16	27.05	41.29	22.1	23.33	24.44
IEMO-IEMO-MSP	19.4	21.18	47.75	37.19	30.77	32.75
MSP*-MSP*-MSP*	53.92	52.67	51.07	55.04	55.98	57.0
IEMO-MSP*-MSP*	53.92	51.47	51.32	55.86	55.42	47.62
IEMO-IEMO-MSP*	41.83	45.57	45.75	47.81	46.22	48.09

Table 5: F1 scores for Anger from experiments using Wav2Vec 2.0, on audio-only and audio+text modalities. Red is the lowest score for each row, yellow the medium score, and green the highest.

speech.

4.3 Multimodal

We focus on Wav2Vec 2.0 as our SER model in our multimodal experiments as it demonstrated marginally better performance in our single-modality experiments. The SER model is pre-trained and fine-tuned on the respective datasets in Table 3; the TER model (BERT) is fine-tuned with the union of the two. Surprisingly, the "PTAPT improvement" column suggests that while PTAPT shows great improvement on IEMOCAP (9.58% increase in UA), often times this method actually performs worse than vanilla fine-tuning, particularly for the MSP dataset (5.34% decrease in UA).

Instead, DAPT appears to be the most robust method due to its highest overall performance across various dataset combinations: DAPT had the highest unweighted accuracy scores for two out of the three cross-corpus dataset configurations testing on the MSP dataset (rows 3, 4), and likewise two out of the three dataset configurations testing on the MSP* dataset (rows 5, 7). In both cases where DAPT did not perform the best (rows 2, 6), FT-only had the highest unweighted accuracy. Although introducing text as an additional modality to DAPT did not help IEMOCAP (row 1 - DAPT: 76.25% UA, PTAPT: 78.6% UA), it greatly improved overall performance across all pre-training methods for MSP and MSP*, particularly the former. This could suggest that for naturalistic data where emotions may be subtle in the audio signal, adding semantic cues via text is essential.

Interestingly, while all multimodal MSP experiments achieved higher UA scores than their MSP* counterparts, macro-F1 scores were lower for MSP than MSP* in all the continued pre-training methods and data configurations.This hints that the majority classes in the imbalanced dataset were identified more often than the minority classes. This is supported by class level evaluations in the "Audio + Text" column of Table 4: neutral has much higher F1 values in MSP experiments than in MSP*.

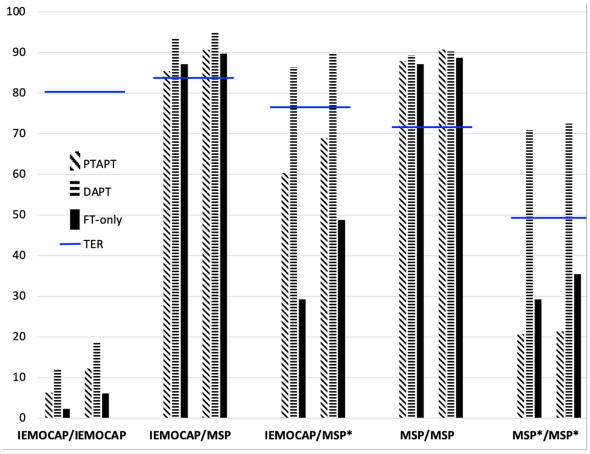

Figure 1: Unweighted Accuracy results for Service Desk experiments. X-axis: experiment setting (pre-training data/fine-tuning data) with audio-only results on the left, audio+transcript on the right. Blue lines are results from the TER models alone.

4.4 Performance on Anger

We are specifically interested in how accurately "anger" is identified. F1 scores for the anger emotion are presented in Table 5. PTAPT generally performs better in both audio and multimodal experiments on IEMOCAP and MSP test sets, but does show some variability in cases where MSP* is both the fine-tuning and test data.

Multimodal results on anger show that adding text generally improves performance going from FT-only to DAPT to PTAPT. Compared to audio-only MSP results, the audio+text modality tends to decrease when pre-trained on IEMOCAP (rows 3-4). Since the TER model is fine-tuned with the union of the SER pre-training and fine-tuning datasets, there is likely a discrepancy in the characteristics of the text and audio for IEMOCAP labeled as anger versus that of MSP, which MSP* mitigates with its smaller data size. Though results indicate that PTAPT could be robust to predicting anger in naturalistic speech if the model is trained on audio only, downstream tasks focusing on anger could potentially benefit from semantic consistency between the text and audio inputs, particularly with the PTAPT method.

4.5 Evaluation on Service Desk Calls

Figure 1 shows evaluating models on real-world service desk calls. Models pre-trained and fine-tuned with IEMOCAP yielded the worst results

		Predictions on anger				Predictions on happy				Predictions on neutral			
		anger	happy	neutral	sad	anger	happy	neutral	sad	anger	happy	neutral	sad
FT-only	IEMO-IEMO		7				8			1	222	60	
	MSP-MSP			7				8			6	277	
	IEMO-MSP			7				8			2	281	
	MSP*-MSP*		1		6		1		7		5	112	166
	IEMO-MSP*			3	4		1	1	6		5	127	151
DAPT	IEMO-IEMO		7				8			1	269	12	1
	MSP-MSP			7				8			2	281	
	IEMO-MSP			7				8				283	
	MSP*-MSP*			6	1			8			2	269	12
	IEMO-MSP*			7				8			3	276	4
PTAPT	IEMO-IEMO		5	1	1		8			3	259	3	18
	MSP-MSP			6	1			8			1	241	41
	IEMO-MSP			7				8			7	276	
	MSP*-MSP*		5	1	1		6	1	1	1	166	90	26
	IEMO-MSP*		2	5			5	3		20	48	212	3

Table 6: Prediction counts for anger, happy, and neutral on Service Desk calls. Total of 7 samples for anger, 8 for happy, 283 for neutral.

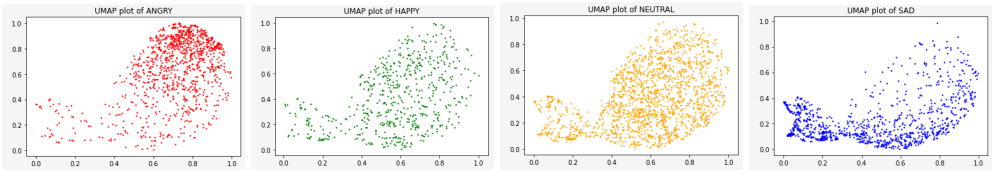

Figure 2: UMAP plots of emotions from the IEMOCAP dataset. From left to right: anger, happy, neutral, sad. Features are Delta and MFCC coefficients from Mel Spectrograms. n_neighbors=200.

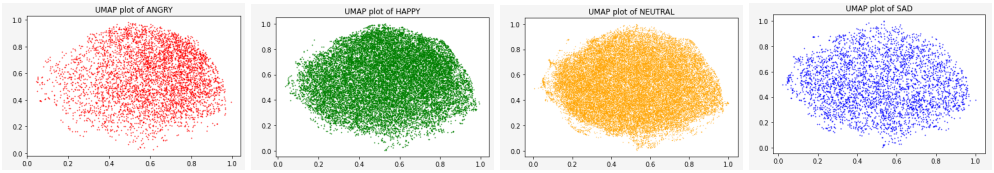

Figure 3: UMAP plots of emotions from the MSP dataset. From left to right: anger, happy, neutral, sad. Features are Delta and MFCC coefficients from Mel Spectrograms. n_neighbors=200.

irrespective of the modalities applied, even lower than the corresponding TER-only run. This aligns with Yoon et al. (2018); Hazarika et al. (2018) in that content ("what is said") is more useful than the audio signal ("how it is said") for recognizing emotions in naturalistic data: words used to express emotions may generalize more effectively from acted speech to naturalistic speech, while audio features, such as an exaggerated tone, are not as easily transferable. In fact, in all data combinations, introducing text modality shows benefit.

Models using MSP* also produced poor results, perhaps due to the imbalanced emotion class present in the real-world data as well as the fewer number of data samples compared to MSP. This is supported by the fact that models fine-tuned using MSP have the best performance, particularly the model pre-trained with IEMOCAP and fine-tuned with MSP, meaning that having more data but tuned to the imbalance seems to yield the best performance. For all the runs except one, DAPT outperformed PTAPT overall, which aligns with our findings in Table 3 and suggests that PTAPT is not the most robust and generalizable continued pre-training method.

Results at the emotion level (see Table 6) provide insight: neutral samples (95% of the calls) were often predicted as happy with the model pre-trained and fine-tuned on IEMOCAP for all three pre-training methods. Anger samples were most often predicted as neutral across all pre-training methods and datasets except when only using IEMOCAP, which consistently predicted happy.

5 Data Visualization

Figures 2 and 3 present the Delta and MFCC Coefficients from Mel Spectrograms for IEMOCAP and MSP respectively.[7] Noticeably, the patterns in Figure 2, while not absolute, do suggest that there is some clustering of anger and sad, while happy and neutral are more dispersed. However, in Figure 3 very little potential clustering emerges. This may explain why the IEMOCAP dataset has higher baselines and why training on IEMOCAP and testing on MSP and MSP* yields better performance: the emotions in IEMOCAP may simply be more easily discernible, meaning features are more easily learned and identified in an unknown audio segment. We see the same difficulty with the service call data (Figure 4), albeit at a much smaller sample size, as the anger and happy are not clustered in any meaningful way and possibly contributes to the inability of models to identify anger utterances in Table 6.

6 Limitations

While our study highlights the discrepancy between widely used acted datasets and real-world speech despite SOTA SER methods, the quality of these datasets should be considered as well. In particular, qualitative analyses on the IEMOCAP dataset by Probol and Mieskes (2023) found that a small subset of the data samples have features that may negatively contribute to inter-annotator agreement and model training. These include short audio lengths (e.g., 1 second), background noise that interferes with the main speaker's intended emotion, and inconsistency between what is said and how it is said. In addition, due to the small size of the service desk calls used as our test dataset, the findings from this pilot study may not be generalizable to other examples of naturalistic speech.

7 Conclusion

We performed a systematic analysis of applying different SER strategies to more naturalistic audio data across experimental setups. We show a noticeable gap in the application of acted speech to more naturalistic audio sources still persists. While we do find that a weighted multimodal approach provides the most potential avenue for model development, the emotion class distributions must still pattern in ways to enable discernible features for an effective SER model. Future research will investigate enhanced feature selection in SOTA multimodal models, combined with methods for handling distributional shifts of emotions between the source and target corpora to further improve performance.

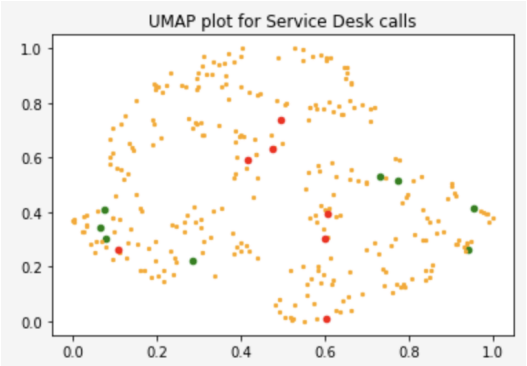

Figure 4: UMAP plots of emotions from the Service Desk dataset. Red: anger. Green: happy. Orange: neutral. Features are Delta and MFCC coefficients from Mel Spectrograms. n_neighbors=200.

Ethics Statement

General speech recognition may bring individual and group privacy concerns. Proper anonymization and containment of personally identifiable information by speech dataset curators is often promised in lab settings, but when data is taken from online platforms like social media or podcasts, this is not necessarily guaranteed. Hence, clarity on how findings from speech recognition research are applied to groups of people is important: conducting audio surveillance in public or obtaining meaningful consent from users of a speech-based product are just some possible examples (Mohammad, 2022). In short, when building real world SER applications, transparency at all levels of the solution is crucial.

Furthermore, SER solutions may inherit or amplify bias from training data used in different stages. For example, an SER system may have higher accuracy in detecting men's emotions than women's; an consequence of this in a medical application may include a delay in detecting warning signals in women patients.

Acknowledgments

This document is export approved by Leidos for release under identification number **24-LEIDOS-1121-28579**.

[7] Visualizations of MSP* show patterns like MSP.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Xuankai Chang, Takashi Maekaku, Pengcheng Guo, Jing Shi, Yen-Ju Lu, Aswin Shanmugam Subramanian, Tianzi Wang, Shu-wen Yang, Yu Tsao, Hung-yi Lee, et al. 2021. An exploration of self-supervised pretrained representations for end-to-end speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 228–235. IEEE.

Li-Wei Chen and Alexander I. Rudnicky. 2021. Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition. *ArXiv*, abs/2110.06309.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 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.

Mustafa Erden and Levent M Arslan. 2011. Automatic detection of anger in human-human call center dialogs. In *Twelfth annual conference of the international speech communication association*.

Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. 2017. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68.

Han Feng, Sei Ueno, and Tatsuya Kawahara. 2020. End-to-end speech emotion recognition combined with acoustic-to-word asr model. In *Interspeech*, pages 501–505.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*.

Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. 2022. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–696, Seattle, United States. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel rahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Reza Lotfian and Carlos Busso. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10:471–483.

Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. 2024. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. *arXiv preprint arXiv:2406.07162*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

Mariana Rodrigues Makiuchi, Kuniaki Uto, and Koichi Shinoda. 2021. Multimodal emotion recognition with high-level speech and text features. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 350–357. IEEE.

Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2015. Anger detection in call center dialogues. In *2015 6th IEEE international conference on cognitive infocommunications (CogInfoCom)*, pages 139–144. IEEE.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Nadine Probol and Margot Mieskes. 2023. Emotions in spoken language-do we need acoustics? In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 71–84.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Trans. Affect. Comput.*, 1(2):119–131.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Florian Eyben, and Björn Schuller. 2022. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:10745–10759.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118.