GENERALIZED ZERO-SHOT AUDIO-TO-INTENT CLASSIFICATION

Veera Raghavendra Elluru, Devang Kulshreshtha, Rohit Paturi, Sravan Bodapati, Srikanth Ronanki

AWS AI Labs

{veerare, kulshrde, paturi, sravanb, ronanks}@amazon.com

ABSTRACT

Spoken language understanding systems using audio-only data are gaining popularity, yet their ability to handle unseen intents remains limited. In this study, we propose a generalized zero-shot audio-to-intent classification framework with only a few sample text sentences per intent. To achieve this, we first train a supervised audio-to-intent classifier by making use of a self-supervised pre-trained model. We then leverage a neural audio synthesizer to create audio embeddings for sample text utterances and perform generalized zero-shot classification on unseen intents using cosine similarity. We also propose a multimodal training strategy that incorporates lexical information into the audio representation to improve zero-shot performance. Our multimodal training approach improves the accuracy of zero-shot intent classification on unseen intents of SLURP by 2.75% and 18.2% for the SLURP and internal goal-oriented dialog datasets, respectively, compared to audio-only training.

Index Terms— audio-to-intent classification, zero-shot, multimodal, teacher-student learning.

1. INTRODUCTION

The use of audio-based systems such as virtual assistants and voice-controlled devices [1, 2], has dramatically increased in popularity over the past few years. The effectiveness of these systems largely depends on their ability to accurately decipher a speaker's intent to provide suitable responses and execute requested actions. Audio-to-intent classification, an essential part of these systems, involves categorizing audio utterances into predetermined intent categories, allowing audio systems to perform specific tasks more efficiently.

Some approaches have been proposed in the past to perform audio-to-intent classification [3, 4, 5]. However, they require large amounts of data, typically in the order of several thousands of utterances, limiting their applicability in the real world. Pre-trained self-supervised learning (SSL) methods using Transformer architecture [6] such as HuBERT [7] and Wav2Vec [8] have shown promising results in various speech-related tasks such as Automatic Speech Recognition [7], Emotion Recognition [9], and Speech Translation [10] with minimal amount of training data for each task. However, their exploration of audio-to-intent classification remains limited [11, 12].

Moreover, spoken language understanding systems that perform audio-to-intent classification [13, 14] assume a fixed set of intent classes during both training and inference. This constraint restricts their applicability in real-world scenarios. Zero-shot Learning [15] can be leveraged for audio-to-intent classification to bridge this limitation. However, despite many works on zero-shot classification in computer vision [16, 17, 18, 19], a limited amount of works exist in the audio processing field. Prior works such as [20, 21] typically propose learning a compatibility function to map audio and class label text in the same embedding space during training, while projecting audio into text space allows classification of audio samples into unseen classes using their textual representations during inference. Another work by [22] follows a similar approach, but additionally proposes a two-way classifier based on a threshold-based decision algorithm to identify if the audio belongs to seen or unseen classes. We also note that the above methods are not specifically for *intent* classification but general audio classification.

The problem of zero-shot classification can be further generalized to predict both seen and unseen intents from the training set. This is often the case in a conversational AI bot system where the full list of intents is not available during training. This scenario is referred to as Generalized Zero-shot Classification (GZSC) [23]. While considerable research has been conducted on GZSC models for textual and image classification [24, 25], there has been a lack of exploration in the realm of audio-based intent classification.

Previous approaches [13, 14] make use of audio-only data for audio-to-intent classification. Various audio classification tasks, such as emotion recognition, have been shown to benefit from multimodal data. For example, a multimodal teacher-student fusion model is proposed in [9] for improving audio-based emotion recognition. Similarly, [26] learns a multimodal speech-text embedding space and uses it for zero-shot audio classification. Inspired by this, we hypothesize that multimodal data can benefit the audio-to-intent classification task as well using audio encoders [7, 8] and pre-trained language models [27, 28]. Additionally, in the context of conversational AI bots, the system is equipped with a set of text sentences provided by the bot developer. These sentences serve as representative examples of each intent that the system can predict. We believe such sentences, if utilized properly, have

the potential to achieve generalized zero-shot audio-to-intent classification. We summarize our contributions as follows:

1. We propose a **multimodal learning** approach for simultaneous training of audio and text encoders using a contrastive loss [29] to align both modalities and a cross-entropy loss for intent classification.

2. We explore conditional **teacher-student learning** approach where the student model mimics the multimodal audio-text representations to improve audio-based intent classifier.

3. We introduce **generalized zero-shot audio-to-intent classification** leveraging a fine-tuned audio encoder from a pre-trained self-supervised model and neural audio synthesizer to create audio embeddings.

2. SUPERVISED AUDIO-TO-INTENT CLASSIFICATION

In this section, we propose to leverage multimodal audio-to-intent classification model to inject lexical information into an audio-based intent classifier. The multimodal classification problem can be more formally introduced as follows: let $A = \{a_1, a_2, .., a_N\}$ be the set of N audio files, $T = \{t_1, t_2, .., t_N\}$ be the set of text transcripts (paired), and $Y = \{y_1, y_2, .., y_N\}$ be the corresponding output intent labels, such that $y_i \in S$. S defines the set of unique intents seen at training time. We can then define a dataset $D = \{a_i, t_i, y_i\}_{i=1}^N$, where N denotes number of examples in a batch. We use a dataset D to train a model f, such that $f(a_i, t_i) = y_i$ for some i.

We describe our methods and proposed approaches in the following subsections, where we leverage pre-trained, self-supervised models to extract audio and text representations, which we use to classify speaker's intent using an end-to-end approach.

2.1. Multimodal learning

Figure 1(a) shows the network architecture of our multimodal model. We use HuBERT [7] as an audio encoder, with which we extract frame-level features from all the utterances in A. We then average the features across each utterance over time and represent them as $\bar{H}_a$. We then project $\bar{H}_a$ onto a lower dimension using linear layer, which we denote as E_a:

$$E_a = W^a * \bar{H}_a + b^a, \tag{1}$$

where W^a and b^a denote the weight matrices and bias of the linear layer, respectively.

We also use BERT [27] as a text encoder, and we convert all the text transcripts in T into a sequence of hidden representations. We average the representations across each transcript over time and represent the result as $\bar{H}_t$. We then project $\bar{H}_t$
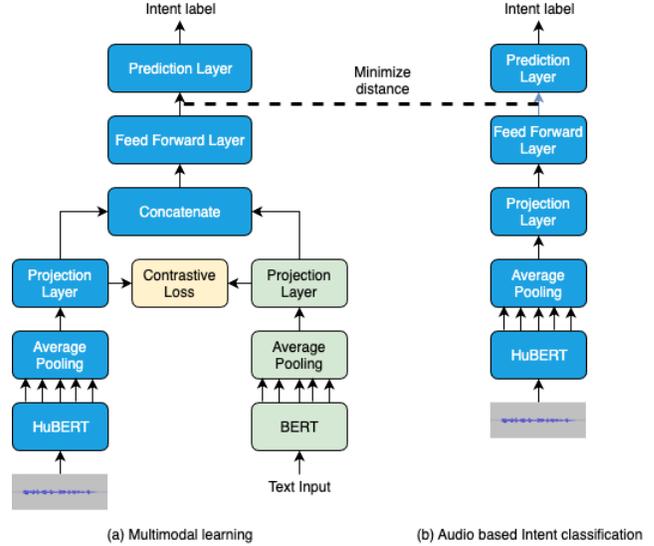

Fig. 1. Architecture of supervised audio-to-intent classification: (a) Multimodal learning using audio and text modalities. (b) Intent classification with audio only modality.

onto a lower dimension using a linear layer, which we denote as E_t:

$$E_t = W^t * \bar{H}_t + b^t, \tag{2}$$

where W^t and b^t denote the weight matrices and bias of the linear layer, respectively.

2.1.1. Intent classification

As shown in figure 1(a), the joint embeddings E_{at} are obtained by concatenating audio representations and text representations followed by a feed forward layer with *ReLu* activation function. These joint embeddings are then passed through a softmax layer to predict the final output $\hat{Y}$:

$$E_{at} = \text{ReLu}(W^{at}(E_a \oplus E_t) + b^{at}), \tag{3}$$

$$\hat{Y} = \text{Softmax}(W^y * E_{at} + b^y), \tag{4}$$

where W^{at}, W^y, b^{at}, b^y denote weight matrices and bias of the concatenation and classification layer, respectively. The model is optimized using cross entropy loss:

$$\mathcal{L}_{IC} = -\frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{S} y_{ij} log(\hat{y}_{ij}), \tag{5}$$

where N is the number of samples in a batch, S is the number of target classes, y_{ij} is the ground truth label for the jth sample and i^{th} class, and $\hat{y}_{ij}$ is the predicted probability for the jth sample and i^{th} class.

2.1.2. Learning a multimodal space

Drawing inspiration from [29, 30], we adopt a joint training approach where we simultaneously train audio and text encoders to acquire a shared multimodal space using contrastive

loss. From Eq. 1 and Eq. 2, the audio ($E_a \in \mathbb{R}^{N \times d}$) and text ($E_t \in \mathbb{R}^{N \times d}$) representations are in the same dimension. We can compare their similarity as:

$$C = \tau * (E_a \cdot E_t^T),$$

where τ is a temperature parameter and the similarity matrix C has N correct pairs in the diagonal and $N^2 - N$ incorrect pairs in the off-diagonal.

The model undergoes training utilizing the contrastive loss [29] paradigm between the audio and text embeddings in pair, following the same loss function in [31]:

$$\mathcal{L}_{CL} = 0.5 * (l_{audio}(C) + l_{text}(C)), \qquad (6)$$

where $l_k = CrossEntropy(C)$ along the text and audio axis, respectively. This leads to the total loss:

$$\mathcal{L}_{MM} = \frac{1}{2}(\mathcal{L}_{IC} + \mathcal{L}_{CL}) \qquad (7)$$

The process of multimodal learning involves training by back-propagating the loss exclusively through the HuBERT model, while keeping the weights of the BERT model fixed. This progressive approach aims to converge the audio embeddings towards the space defined by the text encoder, ultimately leading to the creation of a shared representation space.

2.2. Teacher-student learning

In Section 2.1, we anticipate receiving pairs of audio and text inputs. However, during inference, we lack access to both modalities. Our goal is to predict intent when provided with an audio input. As depicted in Figure 1(b), we exclusively train a model with audio data only. To incorporate lexical information from the joint model, we employ conditional teacher-student learning. This approach leverages transfer learning, where a student model learns to emulate the output distribution of a teacher's model.

The student model is trained using audio data only. The HuBERT model takes in audio signals as input and extracts features at the frame level. We then compute the mean of the frame level features and pass it through a linear layer followed by a feed forward layer with *ReLu* activation function to obtain utterance level audio representation E_s. We then pass E_s through a prediction layer that predicts the intent of the audio. The cross entropy loss computed over predicted intent logits and ground truth (L_{Intent}) is used as a loss function for backpropagation.

By employing transfer learning, we train an audio based intent classification model to generate audio embeddings that closely resemble the joint embedding obtained from the multimodal model. This is achieved by minimizing ℓ_2 norm squared between the two embeddings:

$$\mathcal{L}_{Student} = \|E_{at} - E_s\|_2^2. \qquad (8)$$

The overall training objective of the teacher-student model is then defined as the following weighted combination of the

cross entropy loss and the ℓ_2 norm squared:

$$\mathcal{L}_{Total} = \mathcal{L}_{Intent} + \gamma * \mathcal{L}_{Student}. \qquad (9)$$

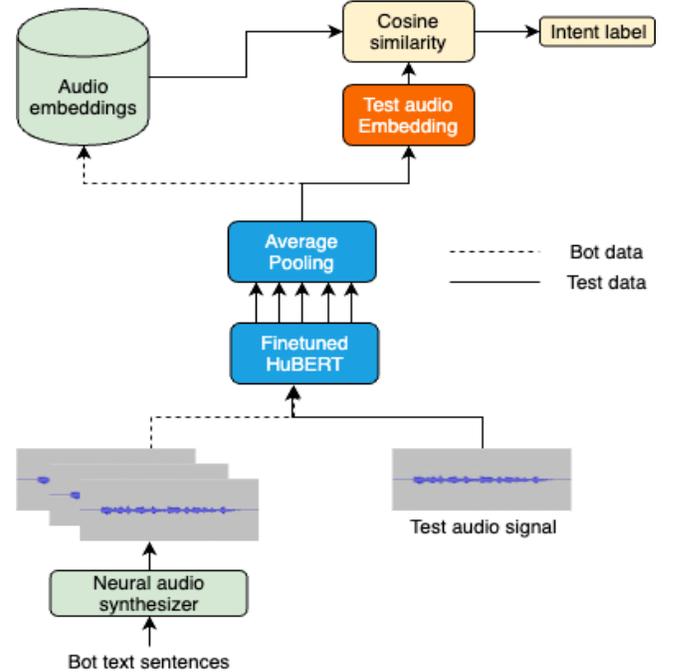## 3. GENERALIZED ZERO-SHOT AUDIO-TO-INTENT CLASSIFICATION

Fig. 2. Generalized zero-shot intent classification utilizing bot text sentences and neural audio synthesizer.

Within a conversational AI bot system, the bot developer provides sample text sentences along with their corresponding intent labels. However, the corresponding audio samples are not provided. It is possible that the intents presented by the bot developer may not align with the intent set used during our training. Consequently, our objective is to predict the intent of an audio utterance that could be either present in the training set or remain unseen during the training process.

3.1. Pre-computation of audio embeddings

To review the generalized zero-shot audio-to-intent classification, we denote $U = \{u_1, u_2, .., u_M\}$ be the set of M text sentences, $V = \{v_1, v_2, .., v_M\}$ be the corresponding intent labels, such that $v_i \in W$, provided by the bot developer. W defines the set of unique intents present in the bot definition. Audio embeddings are created by generating an audio file for each text sentence in U using neural audio synthesizer [32] as shown in Figure 2. Next, we pass each audio file through the HuBERT model and extract frame-level features, followed by average pooling. We associate each audio embedding with its corresponding intent label. This step is conducted only once as an offline process, and all the embeddings are stored in the "audio embeddings" database.

3.2. Inference

Given a test audio file Z, we pass it through the HuBERT model, followed by average pooling. We then compute the cosine similarity between the test audio embeddings and each audio embedding from the embedding database. This similarity calculation helps us determine how closely each audio embedding in the embedding database aligns with the test audio. We identify the utterance in the embedding database that has the highest similarity score to the test audio embedding. We then use the intent label associated with this utterance as the final intent.

4. EXPERIMENTS

4.1. Datasets

Our proposed intent classification approach is trained and evaluated using various datasets, including publicly available SLURP [33] dataset, as well as audio classification datasets such as AudioMNIST [34], TIMIT [35], and Speech Commands [36].

To evaluate zero-shot intent classification, we partition the SLURP dataset, which we refer to as SLURP-Full, into two subsets: SLURP-Seen, which includes 57 intents for the training phase, and SLURP-Unseen, which comprises the remaining unseen 12 intents for the zero-shot evaluation. Additionally, we create another subset, SLURP-mix, consisting of a mix of seen and unseen intents for the generalized zero-shot intent classification task. The SLURP-mix subset consists of 18 intents, of which 12 intents are from SLURP-Unseen and 6 intents are from SLURP-Seen.

Furthermore, to assess the robustness and generalization capabilities of our modeling strategy, we train the teacher and student models on an internal goal-oriented dialog dataset. This dataset consists of 88 hours of audio data and consists of 90 intents from following domains: "Airline", "FastFood", "Finance", "Health-care", "Insurance", "MediaInternetTelecom", "Travel", "Retail", "General". Statistics about train, development and test splits for SLURP and internal intent datasets are provided in Table 1.

Table 1. Datasets used for supervised and zero-shot intent classification

Name	No.Of. Intents	Train	Dev.	Test
Slurp-Full	69	119880	8690	13078
SLURP-Seen	57	111835	7859	12025
SLURP-Unseen	12	8045	832	1053
SLURP-Mix	18	10854	1120	1589
Internal	90	57484	7219	7244

4.2. Models

4.2.1. Baseline models

In our research, we benchmark supervised intent classification task using BERT-CTC model [14], which predicts intent along with ASR by prepending an intent label to the corresponding output sequence. We also evaluate with partially fine-tuned HuBERT model [13], PF-hbt-large, where GRU based decoder is added to decode intent directly from the fine-tuned HuBERT embeddings. For generalized zero-shot intent classification, we consider HuBERT-Frozen as our baseline model. For zero-shot audio classification tasks, we evaluate our model against the semantically aligned speech-text embedding model as the baseline [26].

4.2.2. Multimodal teacher model

The teacher model represents the audio modality with a HuBERT (hubert_large_ll60k) [7] model, the average-pooled audio embeddings of which are projected using a feed forward layer with an output size of 128. Similarly, the text modality is represented using a BERT-base model, the average-pooled embeddings of which are then projected to a 128-dimensional space using a feed forward layer. The multimodal embeddings, obtained by concatenating the audio and text embeddings, are further transformed using a feed forward layer into a 128-dimensional space. Finally, these embeddings are projected to the known dimension of intent space using a softmax layer. We use a dropout of 0.2 and 0.3 for training the unimodal projection layers and multimodal feed forward layer, respectively. Our models are trained using Adam optimizer with an initial learning rate set to 0.0001 and a ReduceLROnPlateau learning rate scheduler. The τ for computing the similarity of audio and text embeddings is set to 0.007. We finetune only the top 6 layers of the HuBERT transformer block. This is because top layers are responsible for capturing language and domain-specific information, while the lower layers capture acoustic information [37]. The BERT model, on the other hand, is frozen.

4.2.3. Student model

Similar to the teacher model, the student model represents the audio modality with a HuBERT model, the average-pooled audio embeddings of which are projected using a feed forward layer with an output size of 128. These projected audio embeddings are further transformed using a feed forward layer and *ReLu* activation function with an output size of 128, which is projected into the known dimension of intent space using a softmax layer. The student model doesn't model any text modality and is trained to minimize the distance between its audio-based embeddings with teacher model's multimodal embeddings as described in Section 2.2. The γ for minimizing the distance between joint embedding representation and audio-only embedding representation is set to 10. Student model also fine-tunes top 6 layers of the HuBERT transformer block.

5. RESULTS AND DISCUSSION

5.1. Supervised Intent Classification

Although the primary focus of this study is zero-shot intent classification, we begin by evaluating the performance of a

Table 2. Supervised Intent Classification accuracy on SLURP-Full and SLURP-Seen datasets. **Top-6**: Top 6 layers of HuBERT transformer block are fine-tuned. **All**: All layers of HuBERT transformer block are fine-tuned.

	SLURP		
Model	Full		Seen
	Top-6	All	
BERT-CTC [14]	87.0	-	-
PF-hbt-large [13]	-	89.22	-
Audio-only	85.43	89.73	84.79
MM	90.20	90.33	91.22
MM-CL	90.77	90.14	91.89
Stu-MM	86.26	**90.43**	86.88
Stu-MM-CL	**87.21**	89.49	**87.46**

multimodal (MM) teacher model and its student counterparts in a supervised intent classification task on the SLURP-Full and SLURP-Seen datasets. We have trained a total of five models and each model is described below.

- **Audio-only**: Model trained using audio-only modality without any guidance from the teacher model.
- **MM**: Model trained using audio and text modalities.
- **MM-CL**: Model trained using audio and text modalities with additional contrastive loss.
- **Stu-MM**: Model trained using audio-only modality with guidance from the MM teacher.
- **Stu-MM-CL**: Model trained using audio-only modality with guidance from the MM-CL teacher.

The results of these evaluations are presented in Table 2. MM and MM-CL models can not be used during inference as we don't have access to the transcript of the audio file. Performance of these models will guide as an upper bound for most of the student and audio-only models. It can be seen that student models (Stu-MM, Stu-MM-CL) outperform baseline BERT-CTC (with K=1) model and audio-only model. The results suggest that student models that are infused with lexical information and trained with contrastive loss are better able to understand the meaning of words, phrases, and the context of an utterance. This improved understanding could help them to perform better on intent classification. To compare our results to those in [13], we also fine-tune all layers of the HuBERT transformer block on the SLURP-Full dataset, as noted in the "All" column. It can be seen that all of our models outperform PF-hbt-large [13] model.

5.2. Generalized Zero-Shot Intent Classification

We evaluate zero-shot intent classification by generating audio embeddings from SLURP-Unseen and SLURP-Mix training sentences, as described in Section 3.1. We then evaluate on the SLURP-Unseen and SLURP-Mix test datasets as explained in Section 3.2. We compare four variants of HuBERT

Table 3. Zero-shot Intent classification accuracy on SLURP-Unseen and SLURP-Mix datasets using models trained on SLURP-Seen and Internal intent datasets

Model trained on	Model Name	SLURP	
		Unseen	Mix
	HuB-Frozen	20.79	24.29
SLURP-Seen	HuB-FT-AO	85.38	87.48
	HuB-FT-MM	87.85	**88.86**
	HuB-FT-MM-CL	**88.13**	88.36
Internal	HuB-FT-AO	28.11	28.13
	HuB-FT-MM	44.25	42.86
	HuB-FT-MM-CL	**46.34**	**45.5**

Table 4. Evaluation of zero-shot intent classification accuracy on AudioMNIST(AM), Speech Commands (SC) and TIMIT

Model	AM	SC	TIMIT
Baseline [26]	15.88	14.27	67.14
HuB-Frozen	**32.06**	25.56	52.86
HuB-FT-AO	16.69	45.87	100
HuB-FT-MM	26.92	**51.26**	100
HuB-FT-MM-CL	29.16	46.85	**100**

models for this task, which are described below. The results are presented in Table 3.

- **HuB-Frozen**: Pre-trained HuBERT model from HuggingFace [38]
- **HuB-FT-AO**: HuBERT model fine-tuned in "Audio-only" model
- **HuB-FT-MM**: HuBERT model fine-tuned in Stu-MM model
- **HuB-FT-MM-CL**: HuBERT model fine-tuned in Stu-MM-CL model

Our analysis indicates that the fine-tuned HuBERT models exhibit a significant performance improvement over the frozen HuBERT model. Furthermore, the student models, HuB-FT-MM and HuB-FT-MM-CL trained using knowledge distillation from the corresponding teacher models outperform the HuB-FT-AO model. This clearly indicates that the lexical infusion and contrastive loss in the teacher model leads to performance improvement.

Table 3 shows that student models trained using teacher-student learning on the mismatched Internal dataset significantly outperform the HuB-Frozen and HuB-FT-AO models. This demonstrates that our training strategy still achieves superior generalization even when trained on mismatched datasets. However, the performance gap between the models trained on SLURP-Seen and Internal datasets emphasizes the importance of training and fine-tuning the models on a representative dataset encompassing different audio and semantic characteristics to achieve generalization.

5.3. Zero-Shot Audio Classification

Given that the audio-based models can be applied to zero-shot audio classification using the generalized zero-shot framework introduced in Section 3, we proceed to assess and compare these models on TIMIT, AudioMNIST and Speech Commands datasets. We created audio embeddings by synthesizing 10 random SX sentences from TIMIT's Test set, digits 0-9 for AudioMNIST, 10 auxiliary words ("Bed", "Bird", "Cat", "Dog", "Happy", "House", "Marvin", "Sheila", "Tree", and "Wow") for Speech Commands. Test sets are used as described in [26]. We then evaluated respective test datasets as discussed in Section 3.2. The results of this evaluation are shown in Table 4. Notably, HuB-FT-MM and HuB-FT-MM-CL models demonstrate significant improvement over the baseline models and HuB-FT-AO. The HuB-Frozen model outperforms all other models on the AudioMNIST dataset, potentially due to its exposure to a substantial number of spoken digits during training. We would like to emphasize that our models achieved accuracy of 100% on the TIMIT dataset. This performance can be attributed to the fact that the linguistic content of the test audio embedding and corresponding audio embeddings in the database are identical.

6. ABLATION STUDY

Validating output from different layers: We conduct an analysis to understand the behaviour when output from the projection layer or feed forward layer is used to calculate the similarity between test audio embedding and each audio embeddings in the embedding database. In Table 5, we present a comparison of the results obtained from these different layers. Our findings reveal an interesting pattern: as the data progresses through the higher layers of the model, it tends to learn task-specific features, leading to a significant drop in accuracy.

Table 5. Zero-shot intent classification using embeddings from Average Pooling layer, projection layer and feed forward (FF) layer.

Model	Average Pooling	Projection Layer	FF layer
HuB-FT-AO	85.37	70.56	65.432
HuB-FT-MM	87.85	76.73	70.84
HuB-FT-MM-CL	88.12	75.499	69.136

Training audio-only model with synthesized data: We develop an audio-only intent classification model by training it on neural audio synthesizer generated audio files using the SLURP-Unseen training and development datasets. The model's architecture remains consistent with the one described in section 4.2.3. This model achieves an accuracy of 88.41% when evaluated on the SLURP-Unseen test data. This result is very close to SLURP-Seen HuB-FT-MM-CL accuracy on SLURP-Unseen test data in the zero-shot intent classification results in Table 3.

Zero-shot performance on different sample sizes: Table 3 demonstrates significantly improved performance when the number of text sentences are in thousands (8045 for SLURP-Unseen). However, it is important to note that in practical scenarios, bot developers may not have access to such an extensive data. Consequently, we conduct an analysis to investigate the performance of our models using few text sentences from each intent, aiming to provide insights into their behavior under more realistic conditions.

We choose 10 to 200 random sentences for each intent from the SLURP-Unseen training set and converted them into audio embeddings as described in Section 3.1 and evaluated on SLURP-Unseen test set with HuB-FT-MM-CL model. We repeated this experiment 10 times to understand how the model's performance varies with different sets of sentences. Figure 3 shows the model's accuracy at different numbers of text sentences per intent. The variation in accuracy is high when the number of sentences per intent is between 10 to 40. Although all models still achieve reasonable accuracy with 10 sentences per intent, the ideal number of sentences per intent is 30, as shown in the graph.
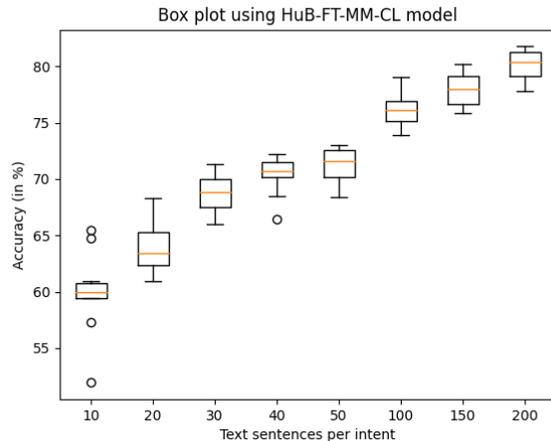

Fig. 3. Accuracy of zero-shot intent classification as size of the text sentences per intent changes

7. CONCLUSION

In this work, we propose training a multimodal model using audio and text modalities and contrastive loss. This model is further used to guide a student model by leveraging transfer learning, which captures the distribution of the joint embedding representation. Our results show that student models achieve higher accuracy than state-of-the-art and audio-only audio-to-intent classifiers. We have reported that generalized zero-shot audio-to-intent classification achieves 2.75% and 18.2% on unseen intents of SLURP using SLURP seen and internal goal-oriented dialog datasets, respectively. For future work, we aim to investigate the clustering of audio embeddings in zero-shot classification and reducing computation overhead during inference.

8. REFERENCES

[1] Gustavo López, Luis Quesada, and Luis A Guerrero, "Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces," in *Advances in Human Factors and Systems Interaction: Proceedings of the AHFE 2017 International Conference on Human Factors and Systems Interaction, July 17- 21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*. Springer, 2018, pp. 241–250.

[2] Cathy Pearl, *Designing voice user interfaces: Principles of conversational experiences*, " O'Reilly Media, Inc.", 2016.

[3] Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters, "From Audio to Semantics: Approaches to end-to-end spoken language understanding," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 720–726.

[4] Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore, "Spoken Language Understanding without Speech Recognition," in *Proc. ICASSP*, 2018, pp. 6189–6193.

[5] Swayambhu Nath Ray, Minhua Wu, Anirudh Raju, Pegah Ghahremani, Raghavendra Bilgi, Milind Rao, Harish Arsikere, Ariya Rastrow, Andreas Stolcke, and Jasha Droppo, "Listen with Intent: Improving Speech Recognition with Audio-to-Intent Front-End," *arXiv preprint arXiv:2105.07071*, 2021.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[8] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[9] Sundararajan Srinivasan, Zhaocheng Huang, and Katrin Kirchhoff, "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," in *Proc. ICASSP*, 2022, pp. 6442–6446.

[10] Ruixiong Zhang, Haiwei Wu, Wubo Li, Dongwei Jiang, Wei Zou, and Xiangang Li, "Transformer based unsupervised pre-training for acoustic representation learning," in *Proc. ICASSP*, 2021, pp. 6933–6937.

[11] Cheng-I Lai, Yung-Sung Chuang, Hung-Yi Lee, Shang-Wen Li, and James Glass, "Semi-supervised spoken language understanding via self-supervised speech and language model pretraining," in *Proc. ICASSP*, 2021, pp. 7468–7472.

[12] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al., "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[13] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba, "A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding," *arXiv preprint arXiv:2111.02735*, 2022.

[14] Yosuke Higuchi, Brian Yan, Siddhant Arora, Tetsuji Ogawa, Tetsunori Kobayashi, and Shinji Watanabe, "BERT Meets CTC: New Formulation of End-to-End Speech Recognition with Pre-trained Masked Language Model," in *Proc. EMNLP*, 2022, pp. 5486–5503.

[15] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell, "Zero-Shot Learning with Semantic Output Codes," in *Advances in neural information processing systems*, 2009, vol. 22.

[16] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov, "DeViSE: A Deep Visual-Semantic Embedding Model," in *Advances in neural information processing systems*, 2013, vol. 26.

[17] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata, "Feature Generating Networks for Zero-Shot Learning," in *Proc. CVPR*, 2018, pp. 5542–5551.

[18] Xiaolong Wang, Yufei Ye, and Abhinav Gupta, "Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs," in *Proc. CVPR*, 2018, pp. 6857–6866.

[19] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang, "Contrastive Embedding for Generalized Zero-Shot Learning," in *Proc. CVPR*, 2021, pp. 2371–2381.

[20] Huang Xie and Tuomas Virtanen, "Zero-Shot Audio Classification via Semantic Embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1233–1242, 2021.

[21] Paul Primus and Gerhard Widmer, "Improved Zero-Shot Audio Tagging & Classification with Patchout Spectrogram Transformers," in *30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 410–413.

[22] Md Tamzeed Islam and Shahriar Nirjon, "SoundSemantics: Exploiting Semantic Knowledge in Text for Embedded Acoustic Event Classification," in *18th International Conference on Information Processing in Sensor Networks*, 2019, pp. 217–228.

[23] Rafael Felix, Ian Reid, Gustavo Carneiro, et al., "Multi-modal Cycle-consistent Generalized Zero-Shot Learning," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 21–37.

[24] AB Siddique, Fuad Jamour, Luxun Xu, and Vagelis Hristidis, "Generalized Zero-shot Intent Detection via Commonsense Knowledge," in *44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1925–1929.

[25] Dwarikanath Mahapatra, Behzad Bozorgtabar, and Zongyuan Ge, "Medical Image Classification Using Generalized Zero Shot Learning," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3344–3353.

[26] Muhammad Huzaifah and Ivan Kukanov, "An Analysis of Semantically-Aligned Speech-Text Embeddings," in *IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 747–754.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019.

[28] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857–16867, 2020.

[29] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "CLAP: Learning Audio Concepts From Natural Language Supervision," in *Proc. ICASSP*, 2023, pp. 1–5.

[30] Soham Deshmukh, Benjamin Elizalde, and Huaming Wang, "Audio Retrieval with WavText5K and CLAP Training," *arXiv preprint arXiv:2209.14275*, 2022.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning Transferable Visual Models From Natural Language Supervision," in *International conference on machine learning*, 2021, pp. 8748–8763.

[32] Amazon Web Services (AWS), "Amazon Polly," 2023, `https://aws.amazon.com/polly/` [Accessed: (6th October 2023)].

[33] Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser, "SLURP: A Spoken Language Understanding Resource Package," in *Proc. EMNLP*, 2020, pp. 7252–7262.

[34] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek, "Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals," *arXiv preprint arXiv:1807.03418*, 2018.

[35] John S Garofolo, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *Linguistic Data Consortium, 1993*, 1993.

[36] Pete Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[37] Pratik Kumar, Vrunda N Sukhadia, and S Umesh, "Investigation of Robustness of Hubert Features from Different Layers to Domain, Accent and Language Variations," in *Proc. ICASSP*, 2022, pp. 6887–6891.

[38] Hugging Face, "Hubert-Large," `https://huggingface.co/facebook/hubert-large-ll60k` [Accessed: (6th October 2023)].