Combining Pretrained Speech and Text Encoders for Continuous Spoken Language Processing

Karan Singla, Mahnoosh Mehrabani, Daniel Pressel, Ryan Price,
Bhargav S. Chinnari, Yeon-Jun Kim, Srinivas Bangalore
karan@whissle.ai, mahnoosh@interactions.com

Abstract

In this paper, we propose a novel architecture for multi-modal speech and text input. We combine pretrained speech and text encoders using multi-headed cross-modal attention and jointly fine-tune on the target problem. The resultant architecture can be used for continuous token-level classification or utterance-level prediction acting on simultaneous text and speech. The resultant encoder efficiently captures both acoustic-prosodic and lexical information. We compare the benefits of multi-headed attention-based fusion for multi-modal utterance-level classification against a simple concatenation of pre-pooled, modality-specific representations. Our model architecture is compact, resource efficient, and can be trained on a single consumer GPU card.

1 Introduction

Speech interfaces have seen wide adoption through virtual assistants such as Siri and Alexa which have rapidly become a part of our everyday lives. To facilitate these applications, high quality automatic systems which infer meaningful information from speech input are essential. This inference is primarily done in two forms, firstly, speech processing (asr, speaker identification, speaker diarization) and secondly, spoken language understanding (text normalization, intent, sentiment). Speech processing applications generally depend on acoustic information derived by functionals or use of pretrained acoustic encoders, however, a typical SLU, the Automatic Speech Recognition (ASR) system is used to convert speech into transcription hypotheses followed by a natural language understanding (NLU) component which acts on those hypotheses to extract an actionable semantic representation. However, in spoken language, organization of acoustic-prosodic cues within an utterance and in-between utterances can resolve semantic, lexical and syntactic ambiguities (Nagel et al., 1996; Snedeker and Trueswell, 2003; Frazier et al., 2006).

Several methods have been proposed to exploit acoustic-prosodic cues along with text for spoken language inference. (Chuang and Wu, 2004; Singla et al., 2018) show that feeding n-best to text classifiers instead of 1-best can boost performance for utterance-level text inference. Some other works show that a speech encoder and text encoders can be jointly optimized for utterance-level multi-modal SLU (Siriwardhana et al., 2020). Alternatively features for speech segments aligned with word embeddings are fed to text based classifier for multi-modal SLU. Recently X combine speech and text encoder using cross-attention between transformer layers of randomly initialized encoder. However, their work is limited to utterance-level SLU (in form of emotion annotations).

In recent work, these encoders are first pretrained on auxiliary tasks either taking speech or text as input. They are then repurposed to further fine-tune using annotations on either modality. We propose to pretrain speech and text encoders before fine-tuning them jointly used supervised data. We show applying one-way cross attention between a text and speech encoder can perform continuous multi-modal tagging of text stream provided by an ASR. As a result, every token in the text accounts for speech variability surrounding it without the need for an explicit alignment. We also show that combining pretrained encoders using two-way cross attention between encoders from multiple modalities shows state of the results for utterance-level emotion and intent prediction.

In this context, pretrained self-supervised encoders, which directly take the continuous input in the form of raw speech, have shown promising results when fine-tuned for transcription tasks. These encoders have also been successfully fine-tuned end-to-end for a variety of SLU tasks (Tzirakis et al., 2017; Chen et al., 2018; Ghannay et al., 2018; Yadav et al., 2020). We start training from a pretrained Wav2vec2 model (Baevski et al., 2020) for converting raw speech segments into fixed-dimensional temporal embeddings. In addition, we use a pretrained text encoder to convert text into token embeddings. We then apply a multi-headed attention between these embeddings in both directions, similar to encoder-decoder attention (Bahdanau et al., 2015). or text-based tagging only one-way cross-attention is applied where text encoder attends to speech encoder for continuous multi-modal tagging.

The contributions of our paper is as follows:

- We illustrate that off-the-shelf pretrained encoders when combined using cross-attention shows state-of-the-art performance (2-6% over text-only models on utterance-level intent and emotion identification.)
- We propose a novel method to attend a pre-trained speech encoder using one-way cross-attention for continuous multi-modal text tagging.
- We show that results for two text token tagging tasks (punctuation insertion in ASR hypothesis and speaker diarization based on ASR hypothesis) improve by 2-4% over text-only model.

2 Related work

We examine self-supervised methods for text and speech encoders, note the rise of using pre-trained speech encoders for superior SLU systems, and touch on the advantages of our method over past multi-modal SLU strategies.

2.1 Pre-trained Speech and Text Encoders

Recently, it has become common practice to first pretrain text encoders using large amounts of unlabeled text before fine-tuning them for a target task (Peters et al., 2017, 2018; Devlin et al., 2018). A popular method of learning text-based, self-supervised encoders is to train a language model to predict the next word in a sequence (Mikolov et al., 2010; Radford and Narasimhan, 2018). BERT (Devlin et al., 2018) introduced a Masked Language Model (MLM) objective, where tokens are randomly masked or perturbed and the model must learn to reconstruct those portions, yielding bidirectional representations. This type of "self-supervision" has also been adopted to encode speech signals (Oord et al., 2018; Pascual et al., 2019; Chung et al., 2019; Baevski et al., 2019). These encoders generally use training targets that are derived from the input signal. For example, the model may be tasked to recover the original input signal given a version transformed through augmentation techniques, recover masked inputs from the future or randomly in the sequence, or separate true inputs from synthetic samples. However, unlike text-based encoders, speech encoders generally need some amount of fine-tuning on a transcription task before being useful for SLU (Chorowski et al., 2015; Chan et al., 2016; Baevski et al., 2020).

2.2 SLU directly from speech

With the emergence of end-to-end ASR (Chorowski et al., 2015; Chan et al., 2016) and the successful pretraining of speech encoders, methods for SLU directly from the speech signal have recently shown comparable performance to the conventional approach of cascading ASR and text-based components in tasks such as named entity recognition (NER), translation, dialogue act prediction (DAP) (Vila et al., 2018; Dang et al., 2020), as well as inference tasks like emotion, intent or behavior understanding (Fayek et al., 2015; Price et al., 2020; Singla et al., 2020).

2.3 Multi-modal SLU

The speech features for multi-modal systems are generally provided either at the level of words or utterances based on the underlying SLU task. Combining speech and text features has led to improved results for multiple tasks including: spoken text parsing, emotion extraction and also for automatic understanding of psychological disorders and human behavior (Yu et al., 2013; Kim and Shin, 2019; Fraser et al., 2013). Unlike previous multi-modal approaches (Kim et al., 2021; You et al., 2021; Tsai et al., 2019), our proposed approach only needs aligned corpora for fine-tuning, not for pre-training. In the past, similar token-level tagging approach has been proposed to spoken text parsing. They perform feature fusion of text and speech features, where speech features are simple functionals representing a word. sequence network for chunk-level multi-modal fusion methods (Tran et al., 2017; Sunkara et al., 2020). Additionally, this is the first work, which performs multimodal token-level classification to improve over speech only and text only approaches for diarization and rich transcription.

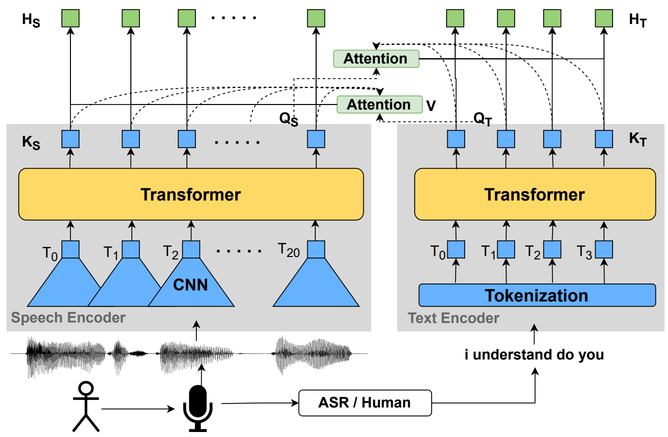

Figure 1: Cross-stitched encoding: Separately pretrained speech and text encoders are combined using a two-way multi-head cross-attention. Output of the attention-level gives token-level speech and text input which has attended to relevant information to decode a token by a supervised fine-tuning task.

3 Cross-stitched Multi-modal Encoder

Cross-stitch[1] is a tiled, raster-like pattern X used repeatedly to form a picture. It has become a common trend to cross-stitch encoders for multi-modal inference (Ye et al., 2019; Wei et al., 2020). (Siriwardhana et al., 2020) apply cross-attention between transformer blocks of randomly initialized speech and text encoders. They then concatenate the pooled output of each encoder for only utterance-level classification, thus not allowing token-level multi-modal fusion. Not performing any pre-training, means, they need a lot more data for fine-training. We propose to combine pretrained speech encoder embeddings with pretrained text encoder embeddings by applying cross-attention on the top of the encoders. We apply two-way multi-headed cross-modal attention between pre-trained speech and text encoders. This allows each encoder to attend to the other modality's encoder in every time-step. Figure 1 gives an overview of the architecture.

The speech and text encoders output K_S and K_T respectively. Keys K_i are either text or speech tokens, and query Q_j is output from the other modality. Following the typical Transformer decoder approach, we first apply self-attention to the target query. Keys and queries are then connected using cross-attention similar to encoder-decoder multi-headed attention (Vaswani et al., 2017). Queries and keys of dimension $[d_q, d_k]$, and values of dimension d_v become inputs to the attention function. We compute the dot products of the query Q_i with all keys K_j and divide each by $\sqrt{d_k}$, where d_{k_j} is dimensionality of keys we are attending. We then apply a softmax function to obtain the weights on the values.

$$Attention(Q_i, K_j, V_j) = softmax(\frac{Q_i * K_{Sj}}{\sqrt{d_{k_j}}}) * V_j$$

We then perform the attention operation h times using different V values where queries, keys and values are low-order projections using W, creating different representations at different positions in the other modality. We employ $h = 8$ parallel attention heads. Multihead cross-attention is formally defined as follows:

$$MultiHead(Q_i, K_j, V_j) = [head_1, .., head_h] * W_j \quad (1)$$

where

$$head_n = Attn.(Q_n * W_Q^n, K_n * W_K^n, V_n * W_V^n) \quad (2)$$

where $W_i^n \in R^{d_{model} \times d_j}$ are parameter matrices. All heads $[1 : h]$ are concatenated to represent each multi-headed token-level cross-attention output for both speech and text input. An additional weight matrix W_j then filters the information from these cross-stitched representations. We use the resultant multi-modal temporal outputs H_S and H_T for various token-level tagging and utterance classification tasks. While performing text tagging, our system only uses H_T and attends to speech encoder via cross-attention. Thus enabling near real-time continuous multimodal SLU. All of our models and experiments are built with an open source library for model exploration and development targeting NLP.

[1] https://en.wikipedia.org/wiki/Cross-stitch

3.1 Speech Encoder

For the speech encoder (SE), we use a Wav2vec2 model with 12 Transformer blocks with 12 attention heads and a 768 dimensional hidden unit size, similar to the base model in (Baevski et al., 2020). Our convolutional feature encoder is adapted for speech data sampled at 8kHz. The model was pretrained on approximately 9450 hours of anonymized speech data from a collection of conversational AI applications where users interact with an intelligent virtual agent (IVA) for customer care over the phone. The model was subsequently fine-tuned with a CTC loss on 900 hours of transcribed data [2].

Our initial testing showed that the lower layers of the architecture contributed most of the information relevant to downstream applications in the multi-modal setting. We found that removing the final 4 Transformer layers from the fine-tuned speech encoder resulted in very little change in performance, but significantly sped up training and inference, while reducing the overall memory footprint. Subsequently, we dropped the final 4 layers of the speech encoder for all experiments.

3.2 Text Encoder

For the text-based encoder (TE), we pretrained an 8-layer Transformer, with 8 attention heads using an MLM loss on a corpus of online data including all of English Wikipedia, around 700 million conversations from Reddit (Al-Rfou et al., 2016; Henderson et al., 2019), 3.3 million online forums, and 8.2 million online reviews for restaurants and hotels. The majority of the dataset contains full conversations between multiple users, and the turns are demarcated with a special end-of-utterance token. Following (Shaw et al., 2018), we use relative positional representations which are not conditioned on the global position of the token but instead use a local relative offset embedding at every layer as part of the self-attention computation. Previous literature has shown that placing the layer norm at the front of each sub-layer in the Transformer simplifies training and can improve performance (Nguyen and Salazar, 2019; Xiong et al., 2020; Wang et al., 2019), so we also follow this approach in our model.

We empirically observed in initial testing that the last 4 layers of the text encoder could be dropped in the downstream multi-modal application without significant performance degradation. As a result, we truncate our text encoder to only the lower 4 of the original 8 layers.

[2]We saw consistent results with publicly available checkpoints.

3.3 Training Details

Our fine-tuning system is compact and lightweight and we are able to train with a single GPU – even on a consumer card. For most experiments, we use a single NVIDIA GTX 1080ti GPU. We use Adam with a fixed batch size of 2 with a fixed learning rate of $1.0e-5$, for all experiments except for IVA intent detection, where we trained with a batch size of 16 on a single A100 GPU [3]. For all experiments, we keep the speech encoder frozen for the first 2000 steps of training. We calculate the cross-entropy loss of a final projection to the number of labels. For tagging, this translates to token-level (*word-level* in our experiments) loss. We use early stopping on a validation set for all experiments.

4 Utterance-level fine-tuning

For spoken utterance classification we compare two fusion methods. First we adopt shallow fusion similar to (Siriwardhana et al., 2020) by first pooling each individual encoder's output (Q_S) for speech and (Q_T) for text. The speech and text pooled output is then concatenated along the embedding dimension. For audio, we use max pooling, and for text, following BERT, we use the special start token ([CLS]). Some datasets contain samples with only text. For these samples, we sum along embedding dimension instead of concatenation to enable smooth training. *SE-TE* refers to shallow fusion and *XSE* refers to the cross-stitched encoder model in Table 5. The unimodal systems using pooling from either Q_S or Q_T.

We train three variants of cross-stitched encoders:

- Pretrained (XSE-P): Pretraining is done as described in section 3 before supervised fine-tuning.
- Scratch (XSE-S): No pretraining is done. Speech and text are encoders initialized from scratch. Cross-attention is applied on both encoders and then output is pooled. This pooled

[3]We used a larger batch size due to the large size of the dataset, to compare against internal benchmarks, and because a grid search yielded significantly better results for that dataset.

outputs are used for joint optimization on supervised corpus.

- Scratch-T (XSE-T): Following (Siriwardhana et al., 2020) we apply cross-attention between speech and text transformer blocks and concatenate the pooled output.

4.1 Emotion Identification

Creating a scalable general purpose solution for emotion extraction comes with the challenge of limited data annotations. *Emotion* which captures behavioral information has been primarily studied in the form of continuous or discrete perceived sentiment (negative, positive, neutral) (Zadeh et al., 2018; Chen et al., 2020), 7 discrete emotions (anger, disgust, fear, joy, sadness, surprise) (Li et al., 2017; Busso et al., 2008) or more granular annotations of behavioral emotion (Demszky et al., 2020). We study emotion as discrete annotations for utterances which have both speech and text available.

Youtube monologues: We report results for widely used CMU-MOSEI (Zadeh et al., 2018) dataset which contains 23,453 annotated video segments from 1,000 distinct speakers and 250 topics, in total approximately 65 hours of speech along with transcriptions. Final sentiment annotated corpora contains 20k sentences annotated by 3 annotators marking discrete sentiment ranging from -3 to 3. We follow the same data setup first as used by (Tsai et al., 2019).

Two-sided telephony conversations: The Switchboard corpus (Godfrey et al., 1992), with 2400 phone conversations from 543 US speakers, was converted to mono-channel audio. Utilizing LDC-provided segmentation, we selected samples with unanimous sentiment labels (positive, negative, neutral), excluding 15% of the data. We allocated 44k segments for training and 2.5k for both development and testing.

Intelligent Virtual Assistant: We also use spoken utterances marked with discrete 7-way sentiment annotated data from an Intelligent Virtual Assistant (IVA) system in the customer care domain. We collect 10K unstructured spoken customer utterances from human-machine dialogue. These utterances/sentences are then coded for sentiment by 3 human annotators, with an agreement of about 75%. We use 8K for training, 1K for development and 1K for testing purposes. We mix data from all annotators for train and test.

Neutral (0) is the dominating label in all datasets, which is also the majority class performance shown in Table 1. Our fusion approaches shallow fusion ($SE-TE$) and cross-stitched fusion (XSE) both outperform text only baselines. XSE performs better than $SE-TE$ for both the MOSEI and IVA dataset. Our shallow fusion system $SE-TE$ is similar to (Siriwardhana et al., 2020) as both concatenate the pooled encoder outputs before classification, however, we use a conversationally-trained, compact MLM instead of the original BERT encoder. On MOSEI dataset (Tsai et al., 2019) report 50.4% vs 53.4% accuracy for our system on 7-way sentiment prediction.

4.2 Intent Detection

Intent detection – attempting to understand a user's goal in a task-oriented dialogue – is a typical problem in SLU. It has primarily been treated as an unstructured prediction problem, applied either independently, or jointly with a separate task to collect specific named entities specific to a conversation (also referred to as slot-filling). For text modality, following (Pressel et al., 2022) we input text encoder a list of the top transcription hypotheses from the ASR system (referred to as N-best lists). We found this yields better results results for text-only system which uses only 1-best provided by an ASR.

We also perform an ablation study to see impact of using additional speech information for few-shot SLU in the form of understanding emotions and intent prediction. We randomly sample N shots for each intent type for training, and use the same development and test datasets. We perform 5 independent runs for both text-based and our multimodal $SE-TE$ and $XSE-P$ setup. Table 2 shows avg. performance across runs.

Intelligent Virtual Assistant We use a large dataset collected from a real-world virtual assistant applications in the customer care domain. It contains approximately 1.1 million anonymized utterances for training. Due to the size of the training set and the cost associated with obtaining human transcription of the spoken utterances and intent labels, N-best hypotheses for the spoken text are taken from a production ASR system consisting of a hybrid DNN-HMM acoustic model and an N-gram language model. Word accuracy of this ASR system is estimated to be in the mid to upper 80% range for this data. The intent labels for training come from two sources. The labels are either generated automatically by an existing production SLU

		Accuracy (%)							
Dataset	Maj.	Speech (SE)		Text (TE)	SE-TE	XSE-P	XSE-S	XSE-T
		no-CTC	with CTC					
MOSEI	40.7	40.9	46.8	46.8	51.7	**53.4**	48.1	50.7
SwitchBoard	48.5	50.2	68.1	69.2	73.3	**73.6**	68.1	70.5
IVA	57	61.2	76.7	79.5	80.2	**80.5**	79.6	80.0

Table 1: Results on emotion identification comparing our text-only approach against proposed multi-modal approaches.

Dataset	Speech (SE)	Text (TE)	SE-TE	XSE-P
IVA	82.34	83.07	84.01	**84.23**
IVA-5shot	28.1	22.4	33.9	**38.1**
IVA-10shot	45.1	50.1	**50.7**	50.4
FSC	99.58	99.34	99.53	**99.63**
FSC-5shot		83.4	86.8	90.2
FSC-10shot		92.3	95.6	97.3

Table 2: Intent detection on IVA and FSC dataset with different modalities.

system when the confidence of the system is very high, or the utterances are sent to a human agent in-the-loop to be manually labeled when the confidence of the automated label is low. The test set consists of approximately 11K utterances that are manually labeled and verified. A development set of approximately 38K noisily annotated utterances is used for early stopping. The dataset has 2 sets of labels indicating intent and entity predictions and, for classification, we use a multi-headed classifier to predict both. The joint accuracy is used to indicate overall performance. For the text modality, the N-best hypotheses are concatenated using a special end of utterance demarcation token (the same end-of-utterance token seen in text pre-training) and passed into the text encoder. While the complete 120 different intent types, for few-shot experiments (5shot, 10shot) we only use 30 most frequent intent types in the training set.

Fluent Speech Commands: We use the publicly available Fluent Speech Commands (FSC) dataset (Lugosch et al., 2019) to train and evaluate our model and compare with models tested on the same dataset. In total, there are 248 different distinct phrases in the FSC dataset and 5 distinct domains. The data are split into 23,132 training samples from 77 speakers, 3,118 validation samples from 10 speakers and 3,793 test samples from 10 speakers. Using human transcriptions our text encoder alone can achieve 100% accuracy. However automatically generated transcripts using ASR are generally noisy. We use the two most likely transcripts generated using an end-to-end ASR model trained with NeMo toolkit. We then use these transcriptions as input to our text encoder.

For the FSC dataset, we observe that, while simple concatenation of the embeddings does not outperform the audio-only encoder, our cross-attention method does better despite a much lower accuracy for the text-only modality (Table 2).

5 Token-level fine-tuning

Our proposed cross-stitched network can be used for multi-modal token-level fine-tuning for both text and speech based classification tasks. In this paper, we focus on doing token-level classification of text tokens where it attends to temporal speech embeddings using multi-headed attention. Figure 2 portrays the multi-modal token-level tagging of text. Rich transcription makes ASR results more readable and valuable for human users. We propose two rich transcription tasks as post-processing on ASR output: 1) Punctuation insertion & capitalization and 2) Speaker diarization in role-based conversations.

5.1 Punctuation insertion & capitalization

We collect 165K English sentences and corresponding speech from Tatoeba[4] to examine if speech aids in punctuation and capitalization. We split this into 141K training, 12K validation, and 13K testing samples. We train our multi-modal system to insert punctuation, specifically, comma (Cm), period (Pr) & question-mark (Qus) and also perform first-letter capitalization (Cp) of words. Our study, focused on the Tatoeba corpus, involves training on normalized text with word-level punctuation and capitalization tags (see sample below).

Input	thank	you	i	understand	do	you
Word tags	Cp:0	0:Pr	Cp:0	0:Pr	Cp:0	0:Qus
Output	Thank	you.	I	understand.	Do	you?

Our system predicts 8 different tags (shown in Table 3) for each word input. Table 3 shows word-level F1-scores for this task and illustrates the improvement in scores using the multi-modal approach (*XSE*) over text-only approach.

[4]https://tatoeba.org/en/

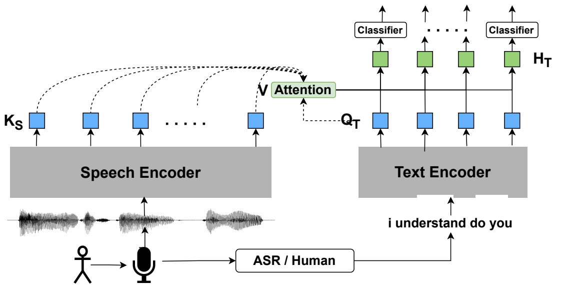

Figure 2: Word-level tagging using cross-attention mechanism. For each word-level prediction in text it takes cross-attn over the corresponding speech segment, thus, doing a soft alignment

Word-level tag		% F1	
		Text	XSE
Punctuation	*Capitalization*		
Comma (,)	Yes	83	**87**
	No	86	**88**
Period (.)	Yes	97	**98**
	No	100	100
Qus (?)	Yes	90	**94**
	No	99	99
None	Yes	100	100
None	No	100	100
Macro-average		93	**95**

Table 3: Results for Punctuation insertion and capitalization task comparing text-only vs proposed multimodal approach (XSE) on Toteba corpus.

5.2 Diarization for role-based conversations

Speaker diarization detects and clusters speaker segments by initially dividing speech into fixed-length frames and then applying hierarchical clustering with a set similarity measure. Alternatively, supervised methods learn to identify speaker changes or apply end-to-end diarization using these frame embeddings.

Our study treats speaker diarization as a task of token-level speaker tagging, focusing on dual-role call-center conversations in the food industry. We use 56 hours of these transcribed and annotated conversations for training, 10 hours for validation, and another 10 hours for testing. The evaluation involves 198 dialogues, encompassing 3.6K speaker turns and 19K words, which our model tags to generate diarized outputs. We hypothesize that because of assigned speaker roles there is a bias between speakers in terms of language use. Below is a sample encoding for two-person role-based conversations.

Mini-batch	A0	A1	A2	C0	C1	C2	C3	C4	A0	A1
	C0	C1	C2	A0	A1	A2	A3	C0	C1	C2
Word tags	1	1	1	0	0	0	0	0	1	1
	0	0	0	1	1	1	1	0	0	0

Here Agent $(A0 - A2)$ words are coded as 1 and client $(C0 - C4)$ words as 0. We train the system to predict 0's and 1's in a continuous stream of words from ASR.

Speaker diarization performance is generally measured using Diarization Error Rate (DER), computed as a sum of false alarms (FA): silence being recognized as speech, missed detections (MD): speech being recognized as silence, and Speaker Error Rate (SER), the % of incorrect speaker tags. In our speech-based results (upper part of Table 4[5]), we report error rates using a typical state-of-the-art speaker diarization approach. We first identify speech and non-speech regions using a Time Delay Neural Network (TDNN) classifier (Bai et al., 2019). Each window of 1.5s length with an overlap of 0.5s is converted into 128-dimensional X-vector (Snyder et al., 2018) by passing through an embedding network trained to classify the speakers of switchboard corpus (Godfrey et al., 1992). We then measure similarity between x-vectors using Probabilistic Linear Discriminant Analysis (PLDA) (Ioffe, 2006; Prince and Elder, 2007). We found using additional unsupervised in-domain corpora (460 hours) translates to improved diarization performance. After measuring the similarity score between all pairs of x-vectors using PLDA, they are clustered until we arrive at two clusters, one for each speaker in the recording. In our work, we found spectral clustering yields better performance than using standard Agglomerative Hierarchical Clustering (AHC) (Lin et al., 2019).

The last two rows of Table 4 shows results for our text-based speaker diarization approach using the cross-stitched encoder which improves over text based role tagging. For token-level word tagging based diarization, we treat word-level error as token error. Our cross-stitched multi-modal approach (*XSE*) shows improvements over text only

[5]We ignore FA errors (at least 6%) as they only account for silence regions in speech.

Approach	% Token error	
	SER+MD	SER
Speech time-series clustering		
VAD + Generic PLDA + AHC	17.1	14.4
VAD + Generic PLDA + Spectral	10.7	7.7
VAD + In-Domain PLDA + Spectral	**7.4**	**4.5**
In-Domain PLDA + Spectral*	5.4	2.9
Token-level role tagging		
Text (TE) - Scratch	16.1	
Text (TE) - Pretrained	8.1	
XSE - Scratch	14.5	
XSE - Pretrained	**7.6**	

Table 4: Token error rates for speaker diarization in 2-person call-center conversations. * is the result with speech vs non-speech segmentations provided by humans.

baseline. Our text based diarization system shows similar performance when compared to a fully automated speech based unsupervised state-of-the-art approach without any in-domain unsupervised data. Best results are achieved for speech-based approach when human provided speech segment information is used instead of automatic voice activity detection (VAD) system.

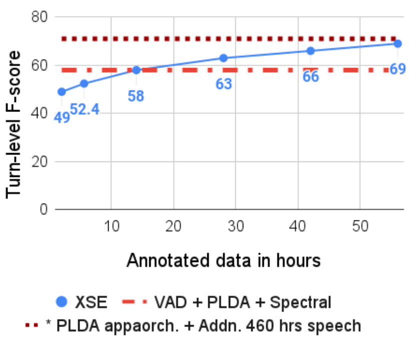

Figure 3: Comparison of multi-modal speaker diarization approach vs typical speech based diarization approach.

Evaluation Metric: Speech-based diarization performs global clustering of speech time frames versus token-level tagging of words which only uses local context. Therefore, we are unable to compare thse approaches directly at the token level. We propose a turn-level evaluation metric for two-person dialogues as high quality transcriptions also implies accurately the whole turn correct. We define Recall (R) as a ratio of number of correct turns to actual turns and Precision (P) is defined as the ratio of number of correct turns to detected turns. F-score is defined as $2PR/(P + R)$ irrespective of length of the segment. Figure 3 shows variation of annotated data (speaker role and boundary information) along with turn-level diarization performance. Figure 3 shows results for multi-modal system using different sizes of annotated corpora. Our proposed approach performs similar to speech-based unsupervised PLDA approach with 14 hrs of annotated corpora. Text-only model shows 65% turn-level F-score compared to 69% for XSE.

Below is a sample output for our cross-stitched embedding (XSE) which takes normalized text and speech as input. It shows combined output of punctuation insertion & capitalization system and also diarization output by performing token-level role tagging.

Input
may i start with your phone number um five one four three eight three remo crescent road nine yes um let's see five one nine four two one uh i don't phone myself so i don't know my damn phone number um five three nine five three nine four one two nine five three nine four one two nine okay so is it pick up or delivery it's a delivery

Output
A: May I start with your phone number?
C: Um five one nine.
A: Yes.
C: Um let's see five one nine four two one. Uh I don't phone myself so I don't know my damn phone number. Um five three nine five three nine.
A: Four one two nine five three nine four one two nine. Okay, so is it pick up or delivery?
C: It's a delivery.

6 Conclusion

Our results show that cross-stitching speech and text encoders using multi-headed attention produces strong results on a diverse set of datasets. Our proposed method supports continuous multi-modal tagging for speech and text input streams. We believe our results can be improved further by including task specific data into unsupervised pre-training of speech and text encoders and exploiting context in dialogue for utterance classification. We plan to explore these directions and evaluate our approach on additional tasks in the future.

We believe our system can be made more robust for near real-time streaming by training with longer sequence lengths and/or by exploiting the context. We plan to extend our approach to more tasks including inverse text normalization, named entity recognition and sentiment tree parsing. Proposed architecture could enable applications such as smart prompting for speech encoders.

References

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *ArXiv*, abs/1606.00372.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Ye Bai, Jiangyan Yi, Jianhua Tao, Zhengqi Wen, and Bin Liu. 2019. Voice activity detection based on time-delay neural networks. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1173–1178. IEEE.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

Eric Chen, Zhiyun Lu, Hao Xu, Liangliang Cao, Yu Zhang, and James Fan. 2020. A large scale speech sentiment corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6549–6555.

Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore. 2018. Spoken language understanding without speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6189–6193. IEEE.

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *arXiv preprint arXiv:1506.07503*.

Ze-Jing Chuang and Chung-Hsien Wu. 2004. Emotion recognition using acoustic features and textual content. In *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, volume 1, pages 53–56. IEEE.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. In *Proc. Interspeech 2019*, pages 146–150.

Viet-Trung Dang, Tianyu Zhao, Sei Ueno, Hirofumi Inaguma, and Tatsuya Kawahara. 2020. End-to-end speech-to-dialog-act recognition. *arXiv preprint arXiv:2004.11419*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. 2015. Towards real-time speech emotion recognition using deep neural networks. In *2015 9th international conference on signal processing and communication systems (ICSPCS)*, pages 1–5. IEEE.

Kathleen C Fraser, Frank Rudzicz, and Elizabeth Rochon. 2013. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *Interspeech*, pages 2177–2181.

Lyn Frazier, Katy Carlson, and Charles Clifton Jr. 2006. Prosodic phrasing is central to language comprehension. *Trends in cognitive sciences*, 10(6):244–249.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699. IEEE.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniel Gerz, Girish Kumar, Nikola Mrksic, Georgios P. Spithourakis, Pei hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. *ArXiv*, abs/1904.06472.

Sergey Ioffe. 2006. Probabilistic linear discriminant analysis. In *European Conference on Computer Vision*, pages 531–542. Springer.

Eesung Kim and Jong Won Shin. 2019. Dnn-based emotion recognition based on bottleneck acoustic features and lexical features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6720–6724. IEEE.

Minjeong Kim, Gyuwan Kim, Sang-Woo Lee, and Jung-Woo Ha. 2021. St-bert: Cross-modal language model pre-training for end-to-end spoken language understanding. In *ICASSP 2021-2021 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7478–7482. IEEE.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Qingjian Lin, Ruiqing Yin, Ming Li, Hervé Bredin, and Claude Barras. 2019. Lstm based similarity measurement with spectral clustering for speaker diarization. *arXiv preprint arXiv:1907.10393*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari.

H Nicholas Nagel, Lewis P Shapiro, Betty Tuller, and Rebecca Nawy. 1996. Prosodic influences on the resolution of temporary ambiguity during on-line sentence processing. *Journal of Psycholinguistic Research*, 25(2):319–344.

Toan Q. Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. *ArXiv*, abs/1910.05895.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Santiago Pascual, Mirco Ravanelli, Joan Serra, Antonio Bonafonte, and Yoshua Bengio. 2019. Learning problem-agnostic speech representations from multiple self-supervised tasks. *arXiv preprint arXiv:1904.03416*.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

Daniel Pressel, Wenshuo Liu, Michael Johnston, and Minhua Chen. 2022. Lightweight transformers for conversational ai. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 221–229.

Ryan Price, Mahnoosh Mehrabani, and Srinivas Bangalore. 2020. Improved end-to-end spoken utterance classification with a self-attention acoustic classifier. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8504–8508. IEEE.

Simon JD Prince and James H Elder. 2007. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training. In *PrePrint*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *NAACL*.

Karan Singla, Zhuohao Chen, David Atkins, and Shrikanth Narayanan. 2020. Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3797–3803.

Karan Singla, Zhuohao Chen, Nikolaos Flemotomos, James Gibson, Dogan Can, David C Atkins, and Shrikanth S Narayanan. 2018. Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. In *Interspeech*, pages 3413–3417.

Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. 2020. Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition. *arXiv preprint arXiv:2008.06682*.

Jesse Snedeker and John Trueswell. 2003. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and language*, 48(1):103–130.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE.

Monica Sunkara, Srikanth Ronanki, Dhanush Bekal, Sravan Bodapati, and Katrin Kirchhoff. 2020. Multimodal semi-supervised learning framework for punctuation prediction in conversational speech. *arXiv preprint arXiv:2008.00702*.

Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. 2017. Joint modeling of text and acoustic-prosodic cues for neural parsing. *arXiv preprint arXiv:1704.07287*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Laura Cross Vila, Carlos Escolano, José AR Fonollosa, and Marta R Costa-Jussa. 2018. End-to-end speech translation with the transformer. In *IberSPEECH*, pages 60–63.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. *ArXiv*, abs/2002.04745.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. *arXiv preprint arXiv:2005.11184*.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021. Self-supervised contrastive cross-modality representation learning for spoken question answering. *arXiv preprint arXiv:2109.03381*.

Zhou Yu, Stefen Scherer, David Devault, Jonathan Gratch, Giota Stratou, Louis-Philippe Morency, and Justine Cassell. 2013. Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs. In *Semdial 2013 DialDam: Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*, pages 160–169.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.