Unsupervised Speech-text word-level alignment with Dynamic Programming

Tianshu Yu[1*], **Zihan Gong**[1,2*], **Guhong Chen**[1,2], **Minghuan Tan**[1†], **Min Yang**[1†],
[1]Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]Southern University of Science and Technology
{ts.yu, zh.gong2, gh.chen2, mh.tan, min.yang}@siat.ac.cn

Abstract

Word-level alignment in speech-text pre-training has demonstrated significant effectiveness, particularly with models like SPECTRA that enhance cross-modal interactions and understand multi-turn dialog contexts. However, these advancements are constrained by a reliance on word-level annotated data, limiting their broader applicability and failing to fully exploit the vast amount of unannotated data available. This paper introduces an **U**nsupervised **S**peech-text word-level alignment with **D**ynamic **P**rogramming (USDP), which reduces the dependency on scarce annotated resources. We propose an iterative training method for USDP, inspired by the EM algorithm. This approach uses Dynamic Programming and EM principles to iteratively refine temporal alignment predictions. Initially, corresponding speech segments are identified based on the model's temporal predictions. A predictor then forecasts text words, and Dynamic Programming is applied to determine the optimal alignment, further refining the model's predictions. Furthermore, we conduct extensive experiments on six benchmark datasets across four different downstream speech-text tasks, including Emotion Recognition in Conversation (ERC), Multimodal Sentiment Analysis (MSA), Spoken Language Understanding (SLU), and Dialogue State Tracking (DST). The experimental results demonstrate that our method significantly enhances the accuracy of models on these speech-text downstream tasks compared to existing approaches.

1 Introduction

In recent years, significant breakthroughs in learning cross-modal feature vector representations have been achieved through speech-text pre-training, which leverages large-scale training corpora to mine data. These successes have been applied across a variety of unimodal and multimodal downstream tasks (Chuang et al., 2019; Kang et al., 2022; Kim et al., 2021b; Li et al., 2021; Tang et al., 2022). These achievements are primarily attributed to the adoption of multimodal self-supervised pre-training loss functions, such as masked modeling and cross-modal contrastive learning, aimed at precisely aligning feature vectors of speech segments with corresponding text sentences.

Despite notable progress in speech-text pre-training models, developing an efficient and integrated model to deeply understand spoken dialogue remains a significant challenge, a problem that has not been sufficiently explored in earlier research. Current methods are mostly designed for specific multimodal downstream tasks, such as speech-to-text translation (Liu et al., 2020) and speech-language understanding (Chung et al., 2020), and struggle to perform as well across a broad range of speech-text tasks as pure text pre-training models do. The word-level alignment-based pre-training methods SPECTRA (Yu et al., 2023) have demonstrated excellent performance across multiple tasks and learned higher quality, finer-grained cross-modal alignments. However, SPECTRA's reliance on word-level alignment annotations not only incurs high costs but is also time-consuming. Therefore, exploring a method that does not require word-level alignment annotations has become the focus of our research.

In the text-to-speech domain, researchers have explored word-level alignment of phonemes in speech and text forms using parallel corpora without word-level alignment annotations. Early methods treat words as sequences of phonemes and use existing aligners, such as teacher models (Ren et al., 2019) or external tools like the Montreal Forced Aligner (Ren et al., 2020), to determine the length of each word or pronounced phoneme. This approach considers the length of each word within the sequence as a random variable, thereby forming

[*] Equal contribution

[†] Min Yang and Minghuan Tan are corresponding authors.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2323–2334
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

a hidden Markov model for the entire sequence of word length variables. Later, Glow-TTS (Kim et al., 2020) introduce a novel algorithm based on auto-encoders and dynamic programming. This algorithm embeds speech and phonemes into the same hidden feature space and then uses dynamic programming to find the most likely matching route. The subsequent developments, such as VITS (Kim et al., 2021a), have improved this method by incorporating optimization techniques like variational auto-encoders and generative adversarial networks, significantly advancing text-to-speech technology. This innovation has propelled VITS to become one of the most popular text-to-speech generation tools on major video platforms today. However, these methods, after converting text into phonemes, essentially discard the semantic information contained in the text. Therefore, they cannot be used for speech-text dialogue pre-training tasks that emphasize understanding.

This paper proposes a novel speech-text dialogue pre-training method based on dynamic programming, which fully utilizes parallel corpora without word-level alignment annotations and significantly reduces the costs of annotation and training compared to previous studies. This method, termed "Unsupervised Speech-text word-level alignment with Dynamic Programming" (USDP), structures each iteration of its training in three phases inspired by the EM algorithm (Moon, 1996), though not strictly following its traditional framework:

- **Estimation:** We use a pre-trained model and a temporal prediction head to estimate the duration of each word.

- **Alignment:** Based on these estimates, we reconstruct the text using corresponding segments of speech. Then, dynamic programming is applied to the prediction matrix—derived from embedding both speech and text into the same hidden space—to determine the optimal alignment path.

- **Optimization:** Finally, we update the model parameters using the predicted alignments and the results of dynamic programming.

Our contributions are summarized as follows:

- We propose a method for word-level alignment pre-training of speech-text using large-scale parallel corpora. This method enhances the effectiveness of speech-text dialogue pre-training models for spoken dialogue understanding, especially in scenarios lacking word-level alignment annotations.

- We address the challenge of word-level alignment in speech-text using dynamic programming, incorporating EM principles for iterative refinement. This approach has yielded positive results in alignment training.

2 Method

In this section, we introduce the data preparation, task of pre-training and the structure of USDP model.

2.1 The Backbone Architecture

We choose SPECTRA (Yu et al., 2023) as our backbone model. It contains three main components: a text-only encoder, a speech-only encoder, and a modality fusion layer, as detailed in Figure 1. During training and application, text is fed into the text encoder and speech into the speech encoder. Then we get the feature vectors of text and speech. The feature vectors are then concatenated and input into the modality fusion layer for cross-modal information fusion. The output from the modality fusion layer, a mixed modality feature vector, is used for pre-training tasks in dialogue understanding or for downstream predictive tasks.

2.1.1 Data Preparation

Before introducing our model and algorithm, we first prepare input text and speech sequence for our model.

We use $D = \{T_1, T_2, \ldots, T_n\}$ to represent a conversation with n dialog turns, Where T_i consists of a slice of raw speech waveform $\mathbf{s}_i$ and it's corresponding text $\mathbf{t}_i = \{\mathbf{w}_{i1}, \mathbf{w}_{i2}, \ldots, \mathbf{w}_{im_i}\}$. Here, $\mathbf{w}_{ij}$ denotes the j-th word of sentence $\mathbf{t}_i$. and m_i denotes the length of sentence $\mathbf{t}_i$. We use s_{ij}/e_{ij} to represent the start/end time of $\mathbf{w}_{ij}$. each word $\mathbf{w}_{ij}$ can be decomposed into a sub-word sequence represented as $\{\mathbf{x}_{ij}^1, \mathbf{x}_{ij}^2, \ldots, \mathbf{x}_{ij}^q\}$($q$ means the length of sub-word sequence).

For each dialog turn T_i, where $i > 1$, we use $k + 1$ ($k \geq 1$) textual dialog turns and 2 speech dialog turns to construct the input $\mathbf{X}_i$, which can be denoted as $\{\mathbf{t}_{i-k}, \ldots, \mathbf{t}_{i-1}, \mathbf{t}_i, \mathbf{s}_{i-1}, \mathbf{s}_i\}$. For efficiency reasons in the pre-training process, we limit the number of speech turns to two, as the length of speech representations typically exceeds that of their corresponding text representations.

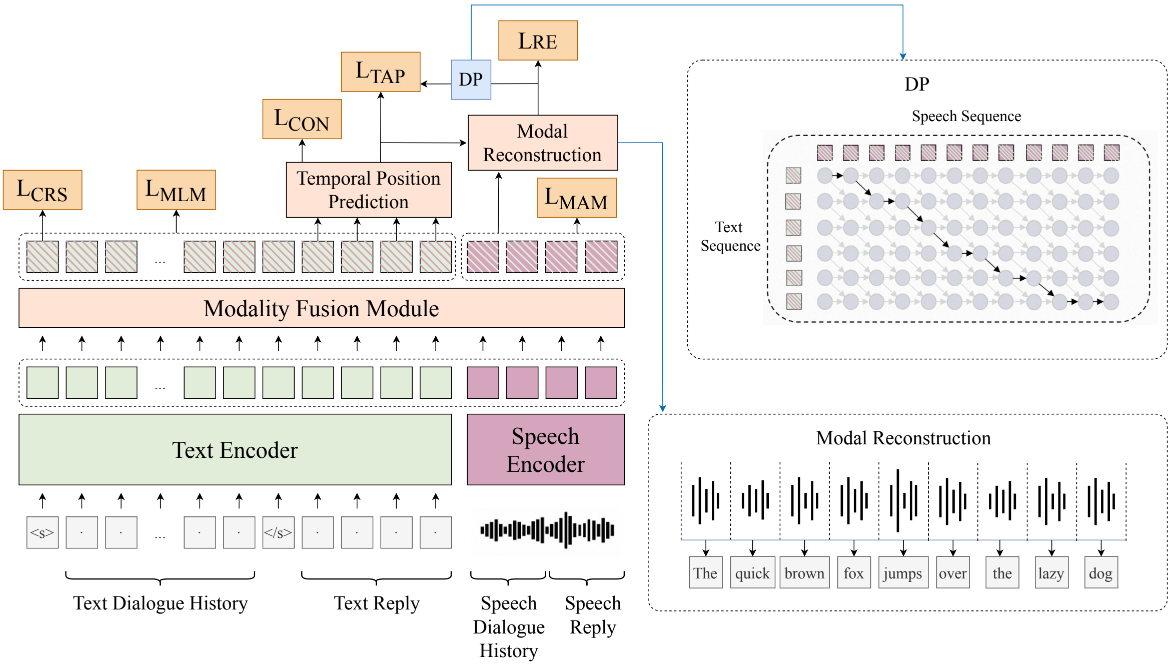

Figure 1: The overview of USDP. The left part shows the overall structure of the pre-trained model. The right part shows the illustration of the process of dynamic programming and modal Reconstruction task.

2.1.2 Model Architecture

Text Encoder Drawing on the exceptional achievements of single-modality pre-trained models in diverse downstream applications, we utilize RoBERTa (Liu et al., 2019) for text encoding. As illustrated in Figure 2 of Appendix B, we use special tokens to delineate sequences and segment dialogues. The <s> token marks the start of a sequence, while </s> signifies the end of each round.

To differentiate dialogs, we introduce learnable segment embeddings $\mathbf{e}_\mathrm{A}$ and $\mathbf{e}_\mathrm{B}$. $\mathbf{e}_\mathrm{B}$ is added to $\mathbf{e}_t^i$ and $\mathbf{e}_{</s>}$, which is the text embedding of $\mathbf{t}_i$ and last token </s>. The sum of text embedding, absolute embedding and segment embedding is then fed into the text encoder, resulting in $\mathbf{H}_i^t$, where $\mathbf{H}_i^t \in \mathbb{R}^{n_t \times d_h}$ denotes the output hidden states of RoBERTa, n_t is the length of textual inputs, and d_h is the dimension of hidden state.

Speech Encoder The speech encoder adopts an architecture similar to that utilized by SPECTRA, which itself is based on the WavLM structure (Chen et al., 2022). As shown by Figure 3 of Appendix C, We add an additional convolutional layer, resulting in each output token of speech features represents approximately 200ms of speech with a stride of 100ms. The parameter of all 8 convolutional layer is shown in Table 4 of Appendix A.

We use the [CLS] and [SEP] tokens to mark the beginning and end of the first speech sequence embedding. The output of the projection layer is then concatenated with the embeddings of the [CLS] and [SEP] tokens. Finally, analogous to the text encoder, segment embeddings ($\mathbf{e}_\mathrm{A}$, $\mathbf{e}_\mathrm{B}$) are added to the concatenated output. The sum of these embeddings is then fed into the self attention layers, resulting in $\mathbf{H}_i^s$, where $\mathbf{H}_i^s \in \mathbb{R}^{n_s \times d_h}$ denotes the output hidden states of last self attention layer, n_s is the length of speech embedding.

Modality Fusion Module We utilize self-transformer attention layers to effectively integrate the two modalities. To differentiate the information from text and speech, we add the corresponding text and speech embeddings (denoted as $\mathbf{e}^t$ and $\mathbf{e}^s$) to the outputs of the text and speech encoders ($\mathbf{H}_i^t$ and $\mathbf{H}_i^s$) respectively. Subsequently, the resulting embeddings are concatenated before being processed by a self-attention layer, yielding the final output $\mathbf{H}_i \in \mathbb{R}^{(n_s+n_t)\times d_h}$.

2.2 Temporal Position Prediction

The model initially takes unmasked text and speech as inputs, and generates temporal predictions using a temporal prediction head. To simplify the issue discussed in this chapter, it is assumed that the first word of the speech starts at 0.0 seconds and continues until the last word ends, with each word closely connected to the next without overlaps or gaps, and with each word lasting a minimum of 0.1 seconds. To ensure the model's temporal alignment predictions remain monotonic, a fully connected layer $\mathbf{W}_o$ is employed to predict the duration of each word.

$$\mathbf{o}_i = \text{Softmax}(\mathbf{H}_i^{(t)}\mathbf{W}_o) \tag{1}$$

where $\mathbf{o}_i \in \mathbb{R}^{m_i \times 1}$ denotes the proportion of each word to the total length of the speech, $\mathbf{H}_i^{(t)} \in \mathbb{R}^{m_i \times h}$ is A matrix composed of the textual hidden representations of the first sub-word of each word.

The length of each word, $\mathbf{l}_i$, can be obtained by multiplying the vector $\mathbf{o}_i$ by L_i, where L_i denotes the duration of the speech for that sentence T_i. The begin time s_{ij} and end time e_{ij} of word $\mathbf{w}_{ij}$ can then be calculated using Equation (2).

$$(s_{ij}, e_{ij}) = (\sum_{k=1}^{j-1} l_{ik}, \sum_{k=1}^{j} l_{ik}) \tag{2}$$

2.3 Pre-training Task

As illustrated in Figure 1, we propose three innovative pre-training objectives for the USDP model, which allow it to effectively train on data devoid of word-level annotations.

2.3.1 Speech to Text Prediction

The purpose of this step is to effectively align and predict the initial sub-word of the textual content that corresponds to the given speech input. This process is crucial for bridging the gap between speech and text representations within the model.

Initially, all text input is masked with [MASK] tokens, ensuring that the model relies solely on the speech input to reconstruct the text. Speech features $\mathbf{H}_i^s$ are processed through a fully connected layer $\mathbf{W}_{st}$, the results $\mathbf{U}_i$ predicts the first sub-word of the corresponding text for each speech feature vector: $\mathbf{U}_i = \mathbf{W}_{st}\mathbf{H}_i^s$ ($\mathbf{W}_{st} \in \mathbb{R}^{V \times d_h}$, $\mathbf{H}_i^s \in \mathbb{R}^{d_h \times L_i}$, where V represents the vocabulary size and d_h the hidden size). The model divides $\mathbf{U}_i$ into different segments according to the temporal predictions (s_{ij}, e_{ij}) and aligns these predictions more closely with the corresponding words. We utilize cross-entropy loss $\mathcal{L}_{\text{RE}}$ to perform a single backward gradient update on $\mathbf{W}_{st}$ while freezing the other parameters of the model.

Algorithm 1 Dynamic programming algorithm for the best alignment under current reconstruction matrix

```
Require: The association matrix $\mathbf{U}_i^*$ of word and speech feature vectors.
Ensure: The length of each word $y_{i1}, y_{i2}, \ldots, y_{im_i}$ is such that the reconstruction loss $\mathcal{L}_{\text{RE}}$ is minimized.
Define $F[1 : L_i, 1 : m_i]$ as the maximum probability of the first $x$ speech feature vectors in $\mathbf{H}_i^s$ corresponding to the first $y$ text word in $t_i$.
$F[1, :] = -\infty$;
$F[:, 1] = \mathbf{U}_i^*[:, 1]$;
for $j = 2$ to $L_i$ do
    for $k = 2$ to $m_i$ do
        $F[j, k] = \max(F[j-1, k-1], F[j-1, k]) + \mathbf{U}_i^*[j, k]$;
    end for
end for
$curr = m_i, last = L_i$;
for $j = L_i - 1; j >= 1; j - -$ do
    if $F[j, curr] \leq F[j, curr - 1]$ then
        $y_{i,curr} = (last - j)/L_i$;
        $curr = curr - 1$;
        $last = j$;
    end if
end for
$y_{i1} = last/L_i$;
```

2.3.2 Optimizing Temporal Alignment Prediction Based on Reconstructed Parameters

After the backward update, the revised matrix $\mathbf{U}_i$ is obtained. From $\mathbf{U}_i$, we extract rows corresponding to the sub-word list $\{x_{i1}^1, x_{i2}^1, \ldots, x_{im_i}^1\}$. The elements of the list represent the first sub-word of words in sentence $\mathbf{t}_i$. We apply softmax to the extracted matrix, resulting in $\mathbf{U}_i^* \in \mathbb{R}^{m_i \times L_i}$, which represents the probability that each audio feature vector in $\mathbf{H}_i^s$ corresponds to each textual word's initial sub-word.

A dynamic programming algorithm is utilized to determine the alignment path that maximizes the total probability, thereby minimizing the loss from speech to text reconstruction. Let $F(x, y)$ represent the highest probability that the first x audio feature vectors in $\mathbf{H}_i^s$ correspond to the first y textual words in t_i. Assuming coherence in speech word sequences, once the word corresponding to

$$F(x,y) = \begin{cases} \mathbf{U}_i^*(x,y) & y = 1, \\ -\infty & x = 1, y \neq 1, \\ \max\{F(x-1,y), F(x-1,y-1)\} + \mathbf{U}_i^*(x,y) & \text{otherwise.} \end{cases} \quad (3)$$

the current audio feature vector is identified, the next audio feature vector must correspond to either the same or the subsequent word. $F(x,y)$ is defined by the dynamic programming state transition equation as shown in Equation (3).

After obtaining $F(L_i, m_i)$ through the state transition equations described earlier, we can trace back the alignment path and identify the selection for each node, thus deriving the feature vector intervals for each word. By measuring the lengths of these intervals, we establish the lengths of each word as $y_{i1}, y_{i2}, \ldots, y_{im_i}$. A schematic of this process is depicted in Figure 1, and the corresponding pseudo-code is outlined in Algorithm 1.

In our experiments, we normalize $\mathbf{y}_i$ by dividing by the speech length L_i, and compute the loss by calculating the KL divergence between $\mathbf{y}_i$ and $\mathbf{o}_i$ as shown in Equation (4):

$$\mathcal{L}_{\text{TAP}} = D_{KL}(\frac{\mathbf{y}_i}{L_i}||\mathbf{o}_i) = \sum_{j=1}^{m_i} \frac{y_{ij}}{L_i} log \frac{y_{ij}}{L_i o_{ij}} \quad (4)$$

2.3.3 Dialogue Consistency Loss

In the experimental data, if $\mathbf{t}_i$ is neither the first nor the last sentence, it is used for temporal predictions in two contexts: $\{\mathbf{t}_{i-k}, \ldots, \mathbf{t}_{i-1}, \mathbf{t}_i\}$ and $\{\mathbf{t}_{i-k+1}, \ldots, \mathbf{t}_i, \mathbf{t}_{i+1}\}$. Enhancing the consistency of temporal outputs across these contexts further constrains the model to produce correct alignments. The consistency loss is computed by calculating the KL divergence between the outputs $\mathbf{o}_i$ from the two predictions:

$$\mathcal{L}_{\text{CON}} = D_{KL}(\mathbf{o}_i^{(1)}||\mathbf{o}_i^{(2)}) = \sum_{j=1}^{m_i} o_{ij}^{(1)} log \frac{o_{ij}^{(1)}}{o_{ij}^{(2)}} \quad (5)$$

where $\mathbf{o}_i^{(1)}$ is temporal prediction result for $\mathbf{t}_i$ in the sequence $\{\mathbf{t}_{i-k}, \ldots, \mathbf{t}_{i-1}, \mathbf{t}_i\}$, and $\mathbf{o}_i^{(2)}$ is temporal prediction result for $\{\mathbf{t}_{i-k+1}, \ldots, \mathbf{t}_i, \mathbf{t}_{i+1}\}$.

2.3.4 Joint Prediction loss

In practical pre-training, while optimizing for temporal prediction, the model is also optimized for cross-modal masked language modeling (CMLM), cross-modal masked acoustic modeling (CMAM) (Li et al., 2021), and Cross-modal Response Selection (CRS) (Yu et al., 2023). The details of CRS task is described in Appendix D These tasks ensure that the model possesses sufficient unimodal understanding capabilities and conversational comprehension skills.

We combine the six pre-training objectives to form a pre-training objective for speech-text pre-training:

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_{\text{RE}} + \mathcal{L}_{\text{TAP}} + \mathcal{L}_{\text{CON}} + \\ & \mathcal{L}_{\text{CRS}} + \mathcal{L}_{\text{CMLM}} + \mathcal{L}_{\text{CMAM}} \end{aligned} \quad (6)$$

3 Experiments

3.1 Pre-training Data

The USDP utilizes pre-training data from the Spotify 100K English podcast dataset (Clifton et al., 2020). The Spotify dataset, drawn from 2020, comprises complete recordings and corresponding text transcriptions of 105,360 audio podcasts, totaling 60,000 hours. Compared to previous works on speech-text pre-training, the dataset used in this paper is more aligned with real-world multi-turn dialogues, making it more suitable for downstream tasks and practical business applications. To facilitate a rigorous comparison with existing speech-text pre-training efforts, we conducts training with a matched dataset size of 960 hours of parallel corpora. Additionally, we limit each audio segment to a maximum of 10 seconds to reduce the impact of prolonged silence, noise, and background music.

3.2 Experimental Setup

Baselines We compare our model with the previous state-of-the-art models (see in Table 1) which are specifically tailored for MSA, ERC, SLU and DST. Additionally, we conduct comparisons with three distinct types of pre-trained models: the text-based RoBERTa (Liu et al., 2019), the speech-oriented WavLM (Chen et al., 2022), and the text-speech multimodal CTAL (Li et al., 2021). Furthermore, our evaluations include comparisons with speech-text multimodal models SPECTRA, which also underwent word-level alignment pre-training but utilized annotated data for training.

Task	MSA				ERC		SLU		DST
Dataset	MOSI		MOSEI		MELD	IEMOCAP	MIntRec		SpokenWoz
Metrics	Acc$_7$	Acc$_2$	Acc$_7$	Acc$_2$	Acc	Acc	Acc$_{20}$	Acc$_2$	JGA
RoBERTa-base	43.75	85.67	51.80	85.88	64.96	64.53	71.24	89.45	20.76
RoBERTa-large	47.81	87.19	54.20	87.23	66.98	66.51	73.71	92.10	21.59
WavLM-base-plus	24.78	65.85	48.92	77.90	32.60	46.90	16.63	55.06	N/D
WavLM-large	33.67	76.52	52.54	81.40	57.77	57.80	57.50	83.10	N/D
CTAL-base	31.34	72.87	50.35	80.36	49.96	53.14	52.80	82.20	13.24
CTAL-large	32.51	72.56	52.60	80.77	52.50	55.12	53.26	81.57	15.79
SPECTRA-base	49.85	87.50	55.33	87.34	68.77	67.94	73.48	91.24	22.05
SPECTRA-large	52.11	88.92	57.29	89.40	**69.65**	69.16	76.17	93.21	**23.44**
SOTA model	MIB		BBFN		M2FNET	M2FNET	MAG-BERT		SPACE
Score	45.70	84.40	51.70	86.10	67.80	66.52	72.13	88.83	20.90
USDP-base	51.01	88.41	56.02	87.97	68.64	67.85	73.93	91.69	22.25
USDP-large	**52.73**	**89.36**	**57.65**	**89.66**	69.52	**69.29**	**76.40**	**93.35**	23.03

Table 1: Comparisons of USDP and other baselines on all downstream tasks. (For better comparison, SPECTRA is not listed in the SOTA model)

Experimental Settings during Pre-training During pre-training, due to the high VRAM consumption of encoding long audio compared to text, each audio segment is limited to a maximum length of 10 seconds. Each segment of audio, along with its corresponding text, is treated as a dialog turn. Additionally, the range of k is set from 1 to 7, meaning that each sample consists of 2 to 8 turns of text and 2 turns of speech.

We conduct two different size models. For USDP, they are called USDP-base and USDP-large. Both text and speech encoder of USDP-base have 12 Transformer layers with a hidden size d_h of 768. Meanwhile, USDP-large has a larger hidden size of 1024, and its text and speech encoder have 24 Transformer layers, respectively. Similarly, for SPECTRA, we have SPECTRA-base and SPECTRA-large. The details of model parameters are shown in Table 4 of Appendix A.

We initialize the text and speech encoders of USDP-base with pre-trained RoBERTa-base and WavLM-base-plus models, respectively. For USDP-large, we use RoBERTa-large and WavLM-large to initialize the text and speech encoders. As our speech encoder includes an additional convolution layer compared to WavLM, only the first seven convolution layers are initialized with pre-trained parameters, and the final layer is initialized randomly.

The USDP model is pre-trained over 100 epochs on eight Nvidia A100 80G GPUs with a batch size of 24 per GPU. Optimization is performed using AdamW (Loshchilov et al., 2017) with a peak learning rate of 1×10^{-4} and a linear warm-up during the initial 1% of updates.

3.3 Model Comparison on Downstream Tasks

Our experimental results highlight the USDP model's significant performance advantages across all four downstream tasks on six datasets, consistently outperforming previous state-of-the-art (SOTA) methods.

3.3.1 Comparison with Previous SOTA model

Fine-tuning on MSA MSA (Hu et al., 2022) aims to predict sentiment labels from multiple modalities. Here , We conduct experiments on two multimodal datasets MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018) to evaluate the effectiveness of our model for the MSA task. We adopt the accuracy over positive/negative sentiments classification (denoted as Acc2) and seven-class classification task (denoted as Acc7) as the evaluation metric for our model and baselines. The experimental results are reported in Table 1.

Specifically, the USDP-large model achieved a Acc7 improvement of 7.03 percentage points on the MOSI dataset over the state-of-the-art (SOTA) MIB method (Mai et al., 2022). It also surpassed the SOTA method BBFN (Han et al., 2021) on the MOSEI dataset by 5.95 percentage points.

Fine-tuning on ERC We also conduct experiments on the ERC task, which requires the model to predict the emotion category of an utterance given a speech clip with its transcripts and dialog history. In this task, USDP-large model outperformed the SOTA method M2FNET (Chudasama et al.,

Settings	TAP	CON	CRS	PTD	MSA		ERC		SLU	DST
					MOSI Acc_2	MOSEI Acc_2	MELD Acc	IEMOCAP Acc	MIntRec Acc_{20}	SpokenWoz JGA
USDP-base	✓	✓	✓	960h	**88.41**	**87.97**	**68.64**	**67.85**	**73.93**	**22.25**
(a)		✓	✓	960h	85.52	86.19	67.05	66.15	71.69	20.87
(b)	✓		✓	960h	88.11	87.55	68.22	67.57	73.48	21.79
(c)	✓	✓		960h	87.96	87.47	67.10	66.28	73.26	20.83

Table 2: Ablation studies of USDP-base. Here, setting (a) (b) (c) means w/o TAP task, w/o CON task and w/o CRS task, respectively.

2022) by 1.72 and 2.77 percentage points on the MELD (Poria et al., 2018) and IEMOCAP (Busso et al., 2008) datasets, respectively, demonstrating a clear enhancement in accuracy.

Fine-tuning on SLU SLU task aims to predict the user intent (Lin and Xu, 2019) given a spoken utterance with the textual transcript. We employ the MIntRec (Zhang et al., 2022) dataset as our experimental dataset for SLU, utilizing 20-class and binary classification accuracy (denoted as Acc20 and Acc2) as evaluation metrics. From Table 1, it is evident that our USDP-large model achieves improvements of 4.27 and 4.52 percentage points in the 20-class results over the previous SOTA method, MAG-BERT (Rahman et al., 2020), respectively.

Fine-tuning on DST In the dialogue state tracking task, we utilize the SpokenWoZ dataset (Si et al., 2023) to evaluate USDP model. The results indicate that our model USDP-large surpasses the performance of previous SOTA method SPACE+WavLM+TripPy (Heck et al., 2020) with an increase of 2.13 percentage points in Joint Goal Accuracy (JGA).

3.3.2 Comparison with Speech/Text model

Compared to models pre-trained on a single modality like RoBERTa and WavLM, USDP demonstrates superior performance in text-related downstream tasks and significant advancements in speech tasks, indicating successful integration of speech and text modalities. The cross-modal alignment training in the pre-training phase is efficient and highly effective. USDP outperforms the previous best speech-text pre-trained model, CTAL, showcasing enhanced performance by precisely aligning cross-modal data and effectively utilizing historical context information. Notably, USDP achieves comparable performance with the SPECTRA model, which utilize word-level alignment annotations. This highlights USDP's efficiency in conducting speech-text word-level alignment training without relying on word-level annotations.

Overall, our method significantly outperforms all comparison models in performance evaluation across six datasets, surpassing even the best prior studies on these datasets. These achievements demonstrate unprecedented effectiveness in key speech-text cross-modal alignment techniques, thanks to our proposed word-level alignment pre-training strategy. The comparison with unimodal pre-training models (such as RoBERTa and WavLM) further confirms the significant benefits of training with both speech and text modalities for enhancing multimodal dialogue understanding. Comparisons with CTAL show that our method can model cross-modal alignment information more effectively. The superior performance on various datasets attests that word-level alignment during pre-training not only achieves higher accuracy but also enhances the model's generalization capability across different speech-text dialogue understanding tasks. The comparison with the SPECTRA models demonstrates that USDP can effectively utilize speech-text parallel corpora without word-level alignment annotations for efficient word-level alignment pre-training.

4 Analysis

4.1 Ablation study

In order to better understand the effectiveness of USDP pre-training method, we investigate the influence of pre-training component on the overall performance of USDP. We report the ablation test results in Table 2.

Impact of TAP Setting (a) "w/o TAP", when compared with the complete USDP-base model, exhibits a significant decline in performance, which further underscores the importance of each component of our model. Specifically, the comparison between setting (a) and the full USDP-base model

sample	right answer	USDP answer	w/o DP answer	text
#1	complaint	✓ complaint	✗ criticism	oh, she is the devil.
#2	notification	✓ notification	✗ suggestion	Please gather in the break room.

Table 3: Intent prediction results on test samples from the MIntRec dataset.

highlights the critical role of word-level annotation pre-training on standard parallel corpora. Appropriate word-level pre-training not only enhances the model's ability to integrate speech and text information but also improves its understanding of the finer-grained connections between the two modalities, which is crucial for boosting overall model performance.

Impact of CON Although the performance of setting (b) does not decline significantly compared to the complete USDP-base model, there still is a noticeable decrease, which indicates that dialogue consistency control plays a positive role in our experimental framework. This is particularly true for tasks that are related to the dialogue context.

Impact of CRS As demonstrated by setting (c), the comparison with the complete USDP-base model further underscores the crucial role of dialogue context pre-training in handling dialogue-related downstream tasks. In setting (c), especially in tasks under multi-turn dialogue scenarios such as ERC and DST, the performance is significantly impacted. Without training with CRS task, the model's ability to utilize dialogue context is notably weakened, thereby affecting task accuracy.

4.2 Case Study

To gain a deeper understanding of the capabilities of our proposed USDP model in learning cross-modal interactions, we carefully select two examples for a detailed case analysis. Both examples are drawn from the MIntRec dataset and are specifically chosen to illustrate scenarios where textual information could lead to confusion. In these cases, integrating the corresponding audio information is essential for a correct interpretation. Building on this, we compare our model with another model from which dynamic programming has been removed (referred to as w/o DP) in ablation studies. This comparison aims to investigate the role of word-level alignment learning in enhancing cross-modal information alignment.

As shown in Table 3, the w/o DP model relies solely on the literal meaning of the text, leading to incorrect predictions of user intent. In contrast, our proposed USDP model, through meticulous word-level alignment training, thoroughly considers the deeper intentions embedded in the speech information, thus enabling accurate prediction of the correct user intentions. These results validate our model's ability to effectively capture and integrate textual and auditory information, subsequently enabling precise prediction of users' true intentions.

On the other hand, w/o DP model exhibits a tendency to overlook or mismanage speech markers, leading to confusion with incorrect labels. This phenomenon reveals a significant flaw in the w/o DP model: the neglect of unique and valuable information within speech data. Furthermore, this underscores the importance of word-level alignment learning in our USDP model. It not only enhances the model's capability to understand cross-modal information but also significantly improves performance in scenarios with ambiguous textual information. This demonstrates the crucial role of precise cross-modal alignment and dialogue understanding in achieving accurate model outcomes.

5 Conclusion

In this paper, we address the coarse granularity of alignment and the inefficient use of conversational context in existing speech-text pre-trained dialogue systems. We propose the USDP method, a novel word-level alignment pre-training approach based on dynamic programming and the Expectation-Maximization (EM) algorithm. Initially, this method utilizes model-outputted temporal alignment predictions to identify corresponding speech segments and uses a predictor to forecast text words. Subsequently, it employs dynamic programming based on the predictor's parameters to find the optimal alignment, thereby refining the temporal alignment predictions of the model. This approach not only resolves the issues of insufficient word-level alignment training and lack of granularity mentioned in SPECTRA but also reduces the dependency on word-level annotations. Experimental results demonstrate that models pre-trained using this method not only surpass single-modality pre-trained models and the best existing speech-text pre-trained models in terms of fine-tuning efficiency on downstream tasks but also significantly outperform the best existing results on these datasets.

Limitation

We analyze the limitations of this work to provide directions for future improvements of our model. Based on our empirical observations, we identify several limitations, which can be divided into two primary categories:

- **Cross-Modal Alignment**: Our proposed method, USDP, is a pre-training approach for speech-text models. However, the field of speech research has yet to develop a counterpart to the image-text alignment achieved by models like CLIP. This gap has hindered progress. Further in-depth research on cross-modal alignment of speech and text using more aligned data, or introducing training techniques from large language models (LLMs) such as contextual learning, could help achieve more effective cross-modal alignment.
- **Dialogue History and Context Understanding**: Although our pre-training method begins to consider dialogue history information, there is still room for improvement in modeling long dialogue histories, understanding contextual nuances, and generating accurate responses. Future research could explore more sophisticated methods for dialogue history modeling and incorporate relevant background knowledge as needed to address the key challenges faced by multimodal dialogue systems.

Acknowledgments

Min Yang was supported by the National Key Research and Development Program of China (2022YFF0902100), the National Natural Science Foundation of China (Grant No. 62376262), the Natural Science Foundation of Guangdong Province of China (2024A1515030166), and Shenzhen Science and Technology Innovation Program (KQTD20190929172835662). Minghuan Tan was supported by the National Natural Science Foundation of China (62406314), China Postdoctoral Science Foundation (2023M733654), Guangdong Basic and Applied Basic Research Foundation (2023A1515110496).

References

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*, 42:335–359.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yung-Sung Chuang, Chi-Liang Liu, Hung-Yi Lee, and Lin-shan Lee. 2019. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. *arXiv preprint arXiv:1910.11559*.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661.

Yu-An Chung, Chenguang Zhu, and Michael Zeng. 2020. Splat: Speech-language joint pre-training for spoken language understanding. *arXiv preprint arXiv:2010.02295*.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, et al. 2020. 100,000 podcasts: A spoken english document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 6–15.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. 2020. Trippy: A triple copy strategy for value independent neural dialog state tracking. *arXiv preprint arXiv:2005.02877*.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Yu Kang, Tianqiao Liu, Hang Li, Yang Hao, and Wenbiao Ding. 2022. Self-supervised audio-and-text pretraining with extremely low-resource parallel data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10875–10883.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021a. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Minjeong Kim, Gyuwan Kim, Sang-Woo Lee, and Jung-Woo Ha. 2021b. 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.

Hang Li, Yu Kang, Tianqiao Liu, Wenbiao Ding, and Zitao Liu. 2021. Ctal: Pre-training cross-modal transformer for audio-and-language representations. *arXiv preprint arXiv:2109.00181*.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. *arXiv preprint arXiv:1906.00434*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov, Frank Hutter, et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*.

Todd K Moon. 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Shuzheng Si, Wentao Ma, Yuchuan Wu, Yinpei Dai, Haoyu Gao, Ting-En Lin, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue in multiple domains. *arXiv preprint arXiv:2305.13040*.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, et al. 2022. Unified speech-text pre-training for speech translation and recognition. *arXiv preprint arXiv:2204.05409*.

Tianshu Yu, Haoyu Gao, Ting-En Lin, Min Yang, Yuchuan Wu, Wentao Ma, Chao Wang, Fei Huang, and Yongbin Li. 2023. Speech-text pre-training for spoken dialog understanding with explicit cross-modal alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7900–7913.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

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.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1688–1697.

Model Type	d_h	Text Encoder	Speech Encoder				Fusion model
		$n_{\text{attn layers}}$	$n_{\text{attn layers}}$	$n_{\text{conv layers}}$	*stride*	*kernel size*	$n_{\text{attn layers}}$
Model-Base	768	12	12	8	[10,3,3,2,2,2,2,5]	[5,2,2,2,2,2,2,5]	1
Model-Large	1024	24	24	8	[10,3,3,2,2,2,2,5]	[5,2,2,2,2,2,2,5]	5

Table 4: Model parameters of two different size of model

A Model Parameters

The model parameters of different size USDP or SPECTRA is shown in Table 4.

B Text Encoder

We split the text into tokens by a BPPE algorithm (Radford et al., 2019) and convert each token into its index according to the dictionary of our tokenizer. Then, we use a pre-trained RoBERTa embedding layer to generate text embedding of each token. The input of text encoder is the sum of segment embedding, absolute positional embedding and the text embedding.

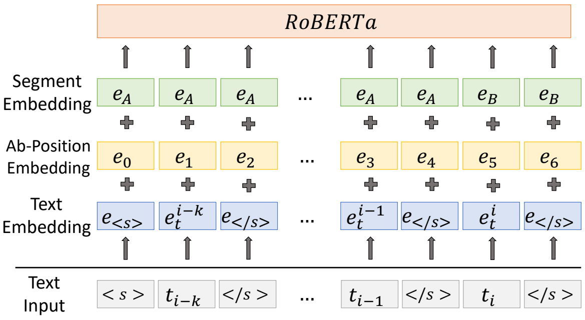
RoBERTa
Segment Embedding
e_A e_A e_A ... e_A e_A e_B e_B
Ab-Position Embedding
e_0 e_1 e_2 ... e_3 e_4 e_5 e_6
Text Embedding
$e_{<s>}$ e_t^{i-k} $e_{</s>}$ ... e_t^{i-1} $e_{</s>}$ e_t^i $e_{</s>}$
Text Input
$<s>$ t_{i-k} $</s>$ ... t_{i-1} $</s>$ t_i $</s>$

Figure 2: RoBERTa input representation. The subscript of absolute position embedding is for illustrative purposes only and does not correspond to actual values.

C The Details Of Speech Encoder

The speech encoder of USDP maintains a similar architecture to the speech encoder of SPECTRA. The encoder comprises of self attention layers, projection layer and several convolutional layers. As detailed in Table 4, the final convolutional layer has a stride of 5, a channel count of 512, and a kernel size of 512, which effectively reduces the length of the extracted speech features. The projection layer consists of layer normalization followed by a fully connected layer that adjusts the dimension of the speech features from 512 to d_h. The self attention layer contains either 12 or 24 WavLM transformer layers.

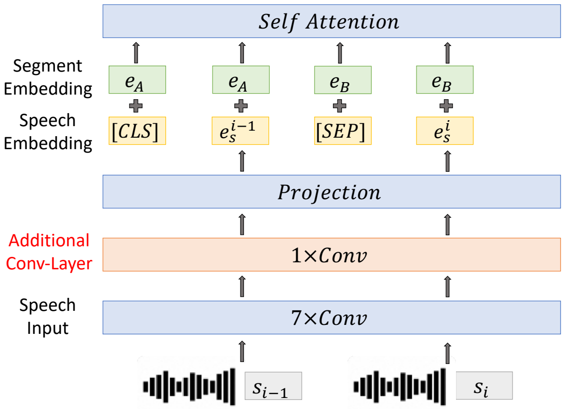

Figure 3: Structure of the speech encoder, modified based on WavLM.

D Details of CRS

For each sample $\mathbf{X}_i$, we randomly replace the text input $\mathbf{t}_i$ or the speech input $\mathbf{s}_i$ with utterances or speech from other dialogues in the dataset. This generates four types of samples for each $\mathbf{X}_i$: (1) only $\mathbf{s}_i$ is randomly substituted; (2) only $\mathbf{t}_i$ is randomly substituted; (3) both $\mathbf{t}_i$ and $\mathbf{s}_i$ are randomly substituted; (4) neither $\mathbf{t}_i$ nor $\mathbf{s}_i$ is substituted. The first three cases are labeled as negative, while the fourth is labeled as positive.

The output of the first <s> token can be viewed as the representation of the whole speech-text sample, so, we apply a softmax function after a fully connected layer on its hidden state to classify the sample into one of the four categories. The cross-modal response selection (CRS) task is optimized using cross-entropy loss, denoted as $\mathcal{L}_{\text{CRS}}$.

E Related Work

SPECTRA SPECTRA is a speech-dialog pre-training model that leverages word-level annotations. Its pre-training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{TPP}} + \mathcal{L}_{\text{CRS}} + \mathcal{L}_{\text{CMLM}} + \mathcal{L}_{\text{CMAM}} \quad (7)$$

where $\mathcal{L}_{\text{TPP}}$ represents the temporal position prediction loss. SPECTRA minimizes the squared error between the predicted word boundaries sij/e_{ij} and their corresponding annotations s^*_{ij}/e^*_{ij}. Details of $\mathcal{L}_{\text{CRS}}$ are provided in Appendix D.

The pre-training objective of USDP partially overlaps with that of SPECTRA, incorporating the same tasks: CMLM, CMAM, and CRS. As indicated by Equations 6 and 7, these tasks play a crucial role in optimizing the model's alignment and contextual representation.