Multimodal Dialog Act Classification for Conversations With Digital Characters

Philine Witzig
Department of Computer Science, ETH Zurich
Zurich, Switzerland
philine.witzig@inf.ethz.ch

Rareș Constantin
Department of Computer Science, ETH Zurich
Zurich, Switzerland
raresionut1@gmail.com

Nikola Kovačević
Department of Computer Science, ETH Zurich
Zurich, Switzerland
nikola.kovacevic@inf.ethz.ch

Rafael Wampfler
Department of Computer Science, ETH Zurich
Zurich, Switzerland
rafael.wampfler@inf.ethz.ch

ABSTRACT

Dialog act classification is essential for enabling digital characters to understand and respond effectively to user intents, leading to more engaging and seamless interactions. Previous research has focused on classifying dialog acts from transcriptions alone due to missing multimodal data. We close this gap by collecting a new multimodal (i.e., text, audio, video) dyadic dialog dataset from 60 participants. Based on our dataset, we developed a novel multimodal Transformer-based dialog act classification model. We show that our model can predict dialog acts in real-time on four classes with a Macro F1 score up to 80.81, outperforming the unimodal baseline by 1.24%. Our analysis shows that the segments of a sentence associated with the highest acoustic energy are most predictive. By harnessing our new multimodal dataset, we pave the way for dynamic, real-time, and contextually rich conversations that enhance the experience of interactions with digital characters.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**; **Neural networks**; **Lexical semantics**.

KEYWORDS

Dialog act classification, Multi-modal learning, Contextual modelling

ACM Reference Format:
Philine Witzig, Rareș Constantin, Nikola Kovačević, and Rafael Wampfler. 2024. Multimodal Dialog Act Classification for Conversations With Digital Characters. In *ACM Conversational User Interfaces 2024 (CUI '24), July 08–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3640794.3665541

CUI '24, July 08–10, 2024, Luxembourg, Luxembourg

ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665541

1 INTRODUCTION

Creating a conversational agent that can mimic human communication is a challenging task due to the inherently multimodal nature of human interaction. The agent should be carefully customized to align with the particular task it is designed to accomplish [5]. This mainly involves providing adequate responses to the user's utterances depending on the given context. Embodied conversational agents (ECAs) should also exhibit social and emotional intelligence through rich verbal and nonverbal behavior [2]. To satisfy these constraints, it is essential to have a strong understanding of the dialog context, which can be achieved by analyzing information from multiple modalities, including text, audio, and visual inputs. One key element in achieving this comprehension is Dialog Act Classification (DAC), a task that is vital for creating a holistic view of the dialog context [28, 29].

Dialog acts are semantic tags that are attributed to utterances with respect to the function they have in the dialog (e.g., "question" or "answer"). Exploiting dialog acts associated with *user* utterances can enhance (non-) verbal response selection for the agent [4, 20]. In turn, the inferred dialog acts corresponding to the *agent*'s utterances can aid the selection of suitable animations for the digital character [10] (see Figure 1). State-of-the-art networks for (DAC) mostly rely on textual information [12, 25, 34, 35, 37, 52, 55]. Only a few works use two modalities (e.g., text and audio) [22, 33]. Nevertheless, employing a multimodal approach that additionally incorporates audio and video could be advantageous for DAC. Prosody can help to better handle ambiguity in syntactically similar utterances [26, 46]. Furthermore, action units in the upper face establish non-verbal signals that can enhance the speaker's intended effect of a dialog act [18].

In this work, we develop a multimodal model for joint DAC on text, audio, and video. Our network consists of three Transformer-based encoders, an inter-modality-attention module, and a sequence model for decoding. We show that our multimodal approach is beneficial for accurately predicting dialog acts in conversations in real-time. Compared to the performance of the unimodal baseline [25], our method achieves an increase of 1.24% in the Macro F1 score on the DailyTalk dataset [30] and 1.7% on a newly collected multimodal dataset called COMOCAP. Our new dataset comprises individual and conversational recordings of 60 participants, leading to a total recording time of 32 hours. Additionally, our method is superior to various other unimodal state-of-the-art approaches

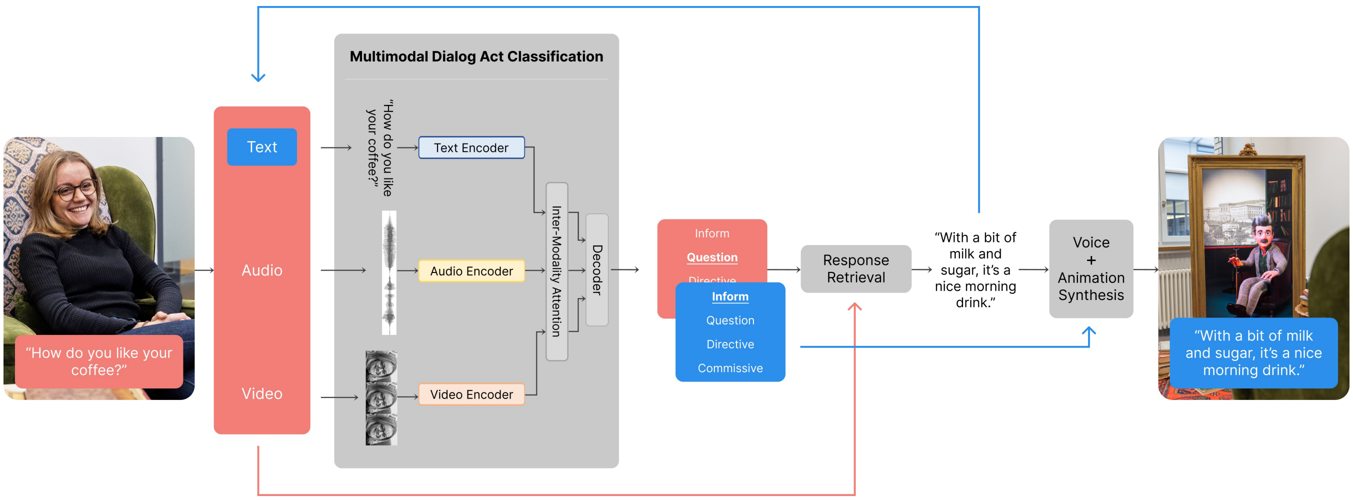


Figure 1: Overview of an exemplary dialog system for conversations between a user and an embodied digital character. We propose a novel Transformer-based model for multimodal dialog act classification. Using all modalities of a user's utterance (red), dialog acts (i.e., inform, question, directive, commissive) are predicted, which can improve character response generation [28, 29]. Similarly, the dialog act predicted from the character's textual response (blue) can be used for improved speech [44] and animation synthesis.

evaluated on DialyDialog [31]. Moreover, we conduct an extensive analysis of the importance of each modality for DAC. We show that the acoustic channel contains predictive features, while the visual channel can be used to stabilize and speed up the learning process of the model. We further demonstrate the applicability of our approach for research on the interaction between users and digital characters by showing that our model is robust to the peculiarities of human-machine conversations. Through the power of our multimodal approach and the versatility of our newly curated dataset COMOCAP, we are advancing the field of dialog act classification laying the groundwork for a new era of dynamic, human-like interactions with digital characters.

1.1 Contributions

Overall, the contribution of our work is fourfold:

- A large-scale, multimodal (i.e., text, audio, video) dataset COMOCAP comprising dyadic dialogs of 60 participants, intended for research in various tasks related to the improvement of the interaction with conversational digital characters.
- A novel multimodal Transformer-based model for real-time dialog act classification based on text, audio, and video, that improves the Macro F1 score on the bimodal dataset DailyTalk [30] by 1.24% and on our new dataset COMOCAP by 1.7%.
- A comprehensive analysis of the importance of each modality for dialog act classification.
- A qualitative and quantitative analysis on the applicability of our method to human-chatbot interactions, providing rationale for integrating our DAC model into ECA dialog systems.

2 RELATED WORK

2.1 Dialog Act Classification

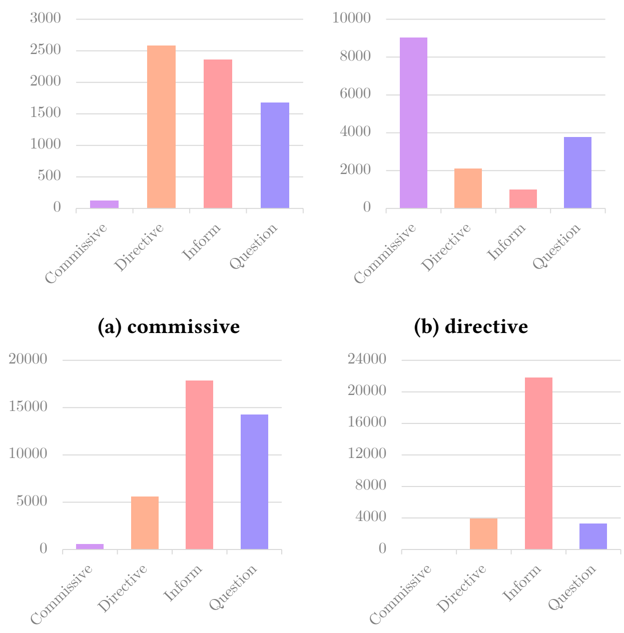
Extensive research has been conducted to enhance the understanding of human communication through dialog acts in various contexts. Most of the existing approaches for DAC exclusively exploit textual features to predict the function of an utterance within a dialog [6, 13, 15, 25, 28, 34, 35, 37]. However, findings in the field indicate that prosodic characteristics, such as pitch and intonation, can be used to modify the intent of an utterance [26, 27]. This suggests that multimodal approaches that make use of both textual and audio inputs could increase DAC performance [22, 33].

Gu et al. [22] and Ortega et al. [33] used CNN-based encoders for text and audio. The features extracted from the different modalities are combined through concatenation. Ortega et al. [33] additionally performed an in-depth analysis of the importance of those modalities. They found that incorporating acoustic features can significantly enhance the model's accuracy, particularly in situations where lexical information (e.g., punctuation) is scarce or absent.

Compared to our work, both works use structurally different datasets, i.e., non-conversational text and audio data from a trauma room [22] and different dialog act taxonomies [33]. However, recent work has shifted the focus back to recognizing intent solely from text [12, 34, 35, 55]. To the best of our knowledge, no research has been conducted to explore the potential benefit of visual cues for DAC. By collecting a conversational, multimodal dataset and training a Transformer-based classifier that can process text, audio, and video, we aim to close this gap in DAC.

2.2 Contextual Modeling

The vast majority of early DAC approaches classified each utterance in the dialog independently [45, 47]. However, the sequential structure of a dialog imposes dependencies between utterances [21]. As a result, the function of the utterance heavily depends on context [38]. This can be supported by looking at the dependencies between dialog acts of consecutive utterances in public datasets, such as DailyDialog [31], which is visualized in Figure 2. The strong inter-tag dependency patterns highlight the importance of contextual information for recognizing intent. For DAC, this implies a notable decrease in the likelihood of a successive dialog act when the dialog act of the preceding utterance is known. Thus, many works proposed a context-based approach for DAC. Bothe et al. [6] included contextual information into their models using an RNN. This increased the final prediction accuracy by 3%. Similarly, Chen et al. [13] proposed CRF-ASN, a hierarchical RNN-based encoder with a conditional random field (CRF) in the decoder. Concurrently, Kumar et al. [28] developed a similar approach based on CRFs. CASA [37] built upon this work by incorporating self-attentive representation learning. Colombo et al. [15] injected guidance into the decoder's attention, improving the accuracy on the SwDa dataset [46] by 4.55%. Due to this strong positive influence on DAC, we only consider contextual modeling.



(a) commissive (b) directive

(c) inform (d) question

Figure 2: Relationship of consecutive dialog acts in DailyDialog [31]. Given the dialog act in the subcaptions a–d, the bar charts illustrate the frequency of occurrence for each dialog act in the next utterance. Since the resulting conditional distributions are not uniform, contextual modeling is essential for DAC.

2.3 Speaker Conditioning

Considering the individuality of each participant in a dialog in terms of language usage, the most recent enhancement in DAC networks involves incorporating speaker information. Shang et al. [42] leveraged speaker information for DAC by integrating it into the layers of their CRF. This increased the performance of their models by 1%. Moreover, He et al. [25] proposed an architecture in which trainable speaker embeddings are fused with the extracted textual features, achieving top performance on multiple benchmarks for DAC. Based on these results, we follow He et al. and incorporate trainable speaker embeddings into our multimodal DAC network.

2.4 Multi-task Learning

In a dialog, intents and emotions constitute valuable knowledge on the conversational context. This led to research in Multi-Task Learning (MTL), where both tags are predicted simultaneously. Cao et al. [9] explored the underlying causal relationships between dialog acts and emotions. They found that emotion can be impacted by the dialog act of the same utterance, while the dialog act can be influenced by the emotion of the same as well as the previous two utterances. Incorporating this knowledge, researchers improved the accuracy of DAC by pushing their networks towards learning emotion-aided embeddings [40]. Since then, numerous multi-task networks have been developed, such as DCR-Net [34] or Co-GAT [35]. Their goal is to reach high performance on both tasks at the same time by including multiple cross-connections between the individual branches of each task. Recent work also took advantage of Large Language Models (LLMs). WEAKDAP used GPT-J to augment multi-task datasets like DailyDialog [31] for further boosting the performance in DAC and Emotion Recognition in Conversations (ERC). Zhao et al. [55] leveraged GPT-3.5 with prompt engineering for DAC and ERC. However, their results indicate that this approach may still be inferior to supervised methods. To primarily maximize performance on DAC, we do not consider MTL in this work and develop a supervised model instead of leveraging an LLM.

3 METHOD

3.1 Preliminaries

A dialog act is a tag that contains information on the semantic and structural function of an utterance in a conversation [41]. It can also be interpreted as the speaker's intent at a lower level [48]. Thus, the terms "dialog act" and "intent" are often used interchangeably. To the best of our knowledge, there is no universally accepted taxonomy for dialog acts. This mainly stems from the numerous factors contributing to the delineation of an utterance's function, including aspects like scope, granularity, and dimensionality.

In this work, we use the dialog act scheme proposed by Amanova et al. [1] consisting of four classes: *question*, *inform*, *commissive* and *directive*. The *question* tag covers all types of questions that can arise in a conversation, while the *inform* tag covers answers, agreements, and disagreements, as well as any other informational statement. The *directive* class represents utterances that are meant to give a direction to the interlocutor in form of a request or a suggestion. Conversely, the *commissive* class captures the commitment or refusal of the speaker to perform a certain action in the present or future. Our choice for this dialog act annotation scheme is mainly driven by two factors. First, it captures general-purpose functions within the dialog. This makes it a suitable taxonomy since we do not target topic- or task-specific conversations. Moreover, it has the benefit that the majority of annotation schemes applied to dialog acts in previously published datasets can be simplified to these four classes.

3.2 Architecture Overview

In this work, we propose a deep multimodal model for DAC, which is visualized in Figure 3. For each utterance, our method processes the three modalities (i.e., text, audio, and video) using individual branches for deep feature extraction. In the text branch, the tokenized sentence is passed through a pre-trained RoBERTa [32] model to obtain a sequence of text embeddings. To extract features for speech representation, we use a pre-trained version of DistilHuBERT [11]. The video branch leverages a pre-trained ResNet-18 [23] backbone followed by a Transformer encoder. To make sure the extracted features align across modalities, we deploy world-level alignment in the audio and video feature extraction branch. Furthermore, we prepend a look-up token $\langle CLS \rangle$, which learns a summarized representation of the entire sequence through the

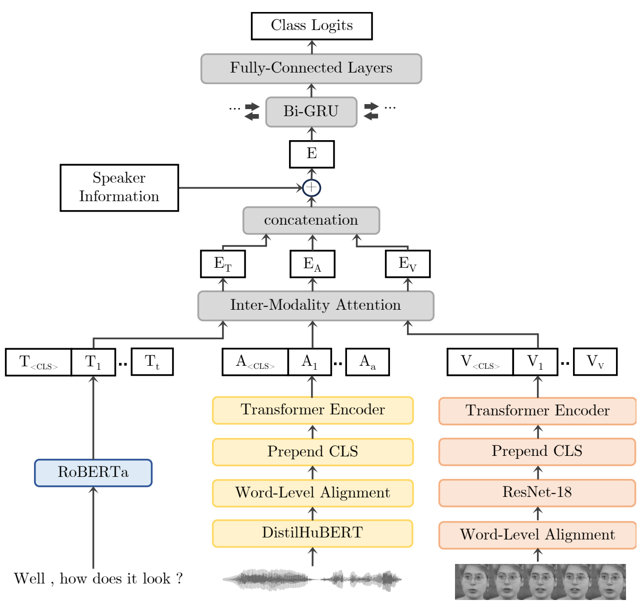


Figure 3: Visualization of our proposed model for multimodal DAC. For each modality (i.e., text, audio, or video) of an utterance, we extract a sequence of deep features including a $\langle CLS \rangle$ token that summarizes information on the entire input sequence. The embeddings are passed through a fusion module that makes use of inter-modality attention. The enhanced classification embeddings are then concatenated and speaker information is added. The final embeddings are fed to a Bi-GRU network. Finally, a set of dense layers is used to obtain a logit for each class.

self-attention mechanism used by the Transformer encoder modules. To encourage the interaction between distinct modalities, their extracted features are passed in pairs through Inter-Modality Attention (IMA) modules [43]. As proposed by He et al. [25], we integrate speaker information into the network by adding trainable embeddings that are conditioned on the tag of the current speaker (i.e., either 0 or 1 since the dialogs are dyadic) to the fused feature vector. Finally, the resulting embeddings of all utterances in the conversation are passed through a Bi-directional Gated Recurrent Unit (Bi-GRU) network. This injects contextual information from the whole dialog into the embeddings. Lastly, the embeddings are passed through a set of fully connected layers to obtain the class logits used for the final prediction. In the following, we elaborate in more detail on the feature extraction branches as well as the IMA module.

3.3 Feature Extraction

3.3.1 Text. Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art method for building rich embeddings from text [17]. BERT embeddings can be used for any task that requires a comprehensive understanding of language. The extracted features capture detailed patterns in input sentences by modeling the embedding of each word and the contextual relationships between them. In the last years, researchers have developed different flavors of BERT for further improving the expressiveness of the text embeddings to generalize to arbitrary downstream tasks. We use one of these versions, which is RoBERTa [32].

Figure 4 illustrates the process of extracting text features from an utterance. In most NLP tasks, it is common to remove the punctuation from the text. However, punctuation is important for intent classification. A single comma or question mark can change the function of the utterance in the dialog. Thus, each punctuation mark must get its own embedding. This is achieved by separating punctuation from words with whitespaces, which are introduced around the punctuation marks. However, this operation is not applied to contractions (e.g., *it's*, *we're*) and abbreviations (e.g., *Mr.*, *p.m.*). They are treated as unitary words in communication. RoBERTa provides

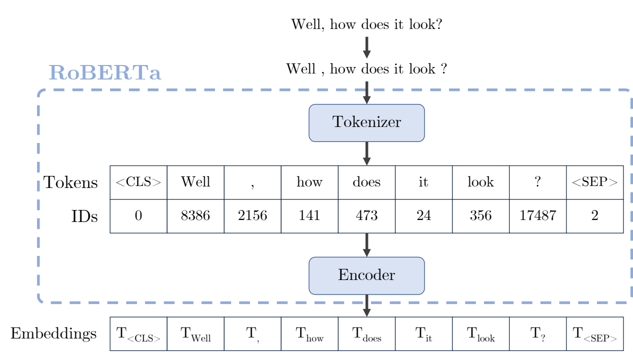


Figure 4: Pipeline for extracting deep features from text. The text is passed through the RoBERTa model, where words are tokenized and each token is embedded into a vector of size 768.

a tokenizer that splits the utterance into tokens. In addition to the contents of the sentence, two additional tokens are appended at the beginning ($\langle CLS \rangle$ token) and end ($\langle SEP \rangle$ token) of the sequence. Finally, text embeddings are obtained by passing the input token identifiers to the RoBERTa encoder generating an embedding vector of size 768 for each token. $T_{\langle CLS \rangle}$, which is the embedding of the $\langle CLS \rangle$ token, acts as a compressed representation of the whole utterance. By fine-tuning the RoBERTa model on a downstream task, the $T_{\langle CLS \rangle}$ embedding can learn expressive sentence-level features that can be used to differentiate between the dialog act classes.

3.3.2 Audio. Our audio feature extraction pipeline is shown in Figure 5. We down-sample the audio signal to 16 kHz and extract the time intervals of each spoken word and pause using the Penn Phonetics Lab Forced Aligner (P2FA) [54]. The pauses are referred to as $\langle SP \rangle$ ("short pause"). We exclude the silent segments at the beginning and end of the actual utterance since they do not provide any useful insights. The truncated audio signal is fed to DistilHuBERT [11] to obtain a temporal sequence of deep features with an embedding size of 768. Each feature vector represents 25 ms of audio. Inspired by Yang et al. [53], we aggregate the features by applying average pooling at the word level using the P2FA alignments. This simplifies the complexity in the network's upper layers, aligning the level of detail with that of the text feature encoder. In addition, we prepend a learnable vector $A^{init}_{\langle CLS \rangle}$ to the embedding sequence before passing it through a Transformer encoder. We empirically found that placing $A^{init}_{\langle CLS \rangle}$ at the beginning of the sequence works best.

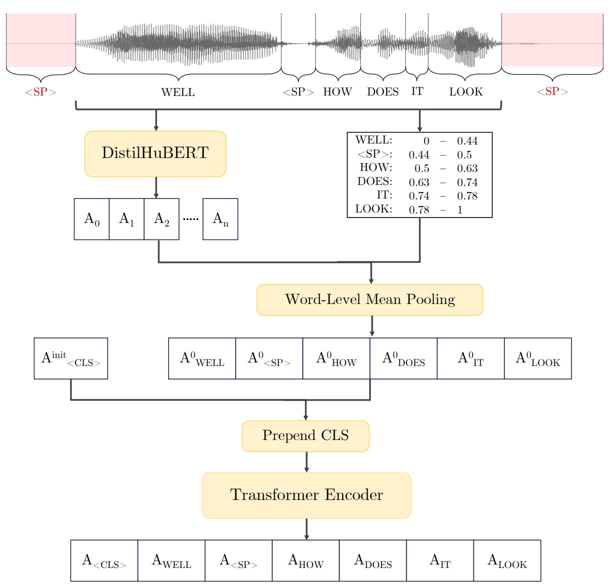


Figure 5: Pipeline for extracting deep features from audio. A forced aligner extracts the start and end timestamps of each spoken word. The audio signal is then passed through DistilHuBERT. The acoustic embeddings are later aggregated using the word alignments and the final sequence of embeddings is obtained after the Transformer encoder.

3.3.3 Video. For extracting video features, we first detect the largest face per frame using the Haar Cascade algorithm [49] and resize it to 48 × 48 pixels in grayscale. Due to efficiency constraints, we select the middle frame for each word using the P2FA alignment timestamps in case a face was detected. The sequence of cropped faces is fed to the pre-trained ResNet-18 [24] to obtain an embedding vector of size 512 for each frame. Similarly to the audio branch, a learnable embedding $V^{init}_{\langle CLS \rangle}$ is prepended to the sequence before passing the input through a Transformer encoder (see Figure 3).

3.4 Inter-Modality Fusion

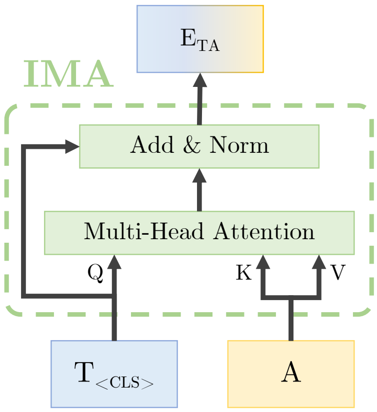
To predict the dialog act of an utterance, we fuse the modality embeddings. Recent approaches, such as CM-Bert [53], utilize cross-modal attention for sharing information across modalities. This requires equal sequence lengths per modality. Although we aggregate the audio and video embeddings at word level, their sequence lengths do not fully match due to punctuation and other additional tokens. Thus, we use the Inter-Modality Attention (IMA) transformer modules proposed by Siriwardhana et al. [43]. In this approach, the $\langle CLS \rangle$ embedding token of each modality is enhanced by the entire embedded sequence of the other modalities.

Figure 6 (left) shows the architecture of the IMA module [43], exemplified for the text $\langle CLS \rangle$ token $T_{\langle CLS \rangle}$ and the audio embedding sequence A. These are passed through a Multi-Head Attention layer, in which the query (Q) corresponds to $T_{\langle CLS \rangle}$, and both the key (K) and value (V) correspond to A. The initial $\langle CLS \rangle$ embedding is added to the output of the attention sub-module through a residual connection. On top, a normalization layer is applied to obtain the embedding of text enhanced through audio features. This is denoted as E_{TA}.

The equations in Figure 6 (right) capture the process of using multiple IMA modules to fuse the features of all three modalities (A for audio, T for text, V for video). First, all pair-wise combinations of modalities are passed through individual IMA modules (Figure 6, Equations 1). Then, for each modality, we perform element-wise multiplication between the two embedding variations and add trainable speaker embeddings [25] (Figure 6, Equations 2). Finally, the enhanced embeddings per modality are fused using concatenation (Figure 6, Equation 3).

3.5 Feature Decoding

Based on the fused multimodal embeddings, the decoder forms a dialog act prediction. It is composed of two modules: the Bi-GRU to leverage contextual information at the conversation level, and the classification head consisting of fully-connected layers. Although Transformers are considered to be the new standard for decoding intra-sequence relationships, it has been shown that simple RNNs suffice to capture the contextual dependencies in dialogs that are relevant when predicting intents [6]. Thus, we use the Bi-GRU [14] to avoid unnecessary complexity in the architecture that could potentially cause overfitting and increase inference time.



$$\begin{aligned} E_{TA} &= \mathrm{IMA}(T_{\langle CLS \rangle}, A) \\ E_{TV} &= \mathrm{IMA}(T_{\langle CLS \rangle}, V) \\ E_{AT} &= \mathrm{IMA}(A_{\langle CLS \rangle}, T) \\ E_{AV} &= \mathrm{IMA}(A_{\langle CLS \rangle}, V) \\ E_{VT} &= \mathrm{IMA}(V_{\langle CLS \rangle}, T) \\ E_{VA} &= \mathrm{IMA}(V_{\langle CLS \rangle}, A) \end{aligned} \tag{1}$$

$$\begin{aligned} E_T &= E_{TA} \odot E_{TV} + S_T \\ E_A &= E_{AT} \odot E_{AV} + S_A \\ E_V &= E_{VT} \odot E_{VA} + S_V \end{aligned} \tag{2}$$

$$E = concat(E_T, E_A, E_V) \tag{3}$$

Figure 6: Visualization of feature fusion using IMA [43]. The diagram of the IMA transformer (left) is shown exemplary for text and audio. In the equations for computing the final embedding E (right), $\odot$ is the element-wise multiplication, and $S_{\mathbf{modality}}$ are learnable speaker embeddings.

3.6 Training

Our model optimizes the Cross-Entropy (CE), i.e., the loss function is

$$\mathcal{L}_{CE}(\hat{z}, y) = -\sum_{c=1}^{C} y_c \log(\hat{z}_c) \tag{4}$$

where y_c is the binary value indicating whether the class c corresponds to the ground truth label or not, while $\hat{z}_c$ represents the predicted logit for class c.

For the pre-trained RoBERTa, DistilHuBERT, and ResNet-18, we only fine-tune the last layer and keep the others frozen during training. The Transformer encoders consist of one layer with four attention heads each. The Bi-GRU comprises two layers. During training, we use AdamW with a weight decay of $5e-4$, a batch size of 1, a chunk size (maximum number of utterances in a dialog) of 20, and 10 epochs for early stopping. We add multiple dropout ($p = 0.5$) and sequence normalization layers throughout the network. Lastly, we make use of a cascade training technique, where the model is first pre-trained on a text corpus. Afterwards, we freeze the text branch and integrate the audio branch for additional training on a text-audio dataset. Finally, we fine-tune the video branch on a text-audio-video dataset. Since this dataset contains novel speakers, we train the parameters of the audio and video branches. At each stage of the fusion process, the decoding layers remain trainable.

4 DATASETS

We train and evaluate our model on three different datasets: DailyDialog [31], DailyTalk [30], and a newly collected dataset (COMOCAP) introduced in Section 4.1. The splits per dataset are shown in Table 1. For DailyDialog, we use the split proposed by Wen et al. [51] since it avoids the overlapping issue where samples from the train and test set overlap. DailyTalk and COMOCAP (scripted dialogs) are subsets of DailyDialog. Hence, their splits should be aligned. Instances from the training set of one dataset should not be included in the test set of the other dataset. Furthermore, DailyDialog and DailyTalk are imbalanced in terms of their label distribution. For DailyDialog, this phenomenon becomes apparent when aggregating the total count of next dialog acts for each dialog act in Figure 2. To compensate for this imbalance, we assign weights to conversations during training. We use a weighted sampler so that the sampled set in each epoch has an almost uniform distribution of labels. The weight w_i of each utterance is inversely proportional to the fraction of the associated label L within the entire corpus R. With l_k being the label of utterance u_k, we obtain:

$$w_i = \frac{|R|}{\left|\left\{u_k \mid u_k \in R, l_k = L\right\}\right|}. \tag{5}$$

The weight W_j of the dialog D_j is the mean of the weights of its utterances u_i:

$$W_j = \frac{1}{|D_j|} \sum_{i=1}^{|D_j|} w_i, \quad \forall u_i \in D_j. \tag{6}$$

Table 1: Train-validation-test split for the three datasets used during training and evaluation, i.e., DailyDialog, DailyTalk, and COMOCAP. We only use the scripted dialogs from COMOCAP. For each dataset, the number of used conversations and utterances in the train, test, and validation set is shown.

	Conversations			Utterances		
	Train	Val	Test	Train	Val	Test
DailyDialog [31]	8,967	1,001	1,004	69,185	7,650	7,999
DailyTalk [30]	2,088	216	233	19,416	1,953	2,337
COMOCAP	606	86	76	2,438	320	304

4.1 COnversational MOtion CAPture Dataset (COMOCAP)

We collected a large-scale, multimodal COnversational MOtion CApture dataset (COMOCAP) of 60 participants. COMOCAP is intended for research in various tasks related to the improvement of conversational digital characters. We built a custom motion capture setup that records audio, video, and animation parameters of participants in dyadic conversations. The recorded sequences can be separated into three categories: (i) scripted sentences, (ii) scripted dialogs, and (iii) free dialogs. Please note that we only use samples from the scripted dialogs in this work.

4.1.1 Participants. We recruited 60 participants (29 identified as female, 31 as male, 0 as non-binary and 0 preferred not to answer) between the ages of 18 and 42 (mean = $24.7\bar{3}$ years, SD = 4.95 years). 12 participants had a background in acting. In Figure 7, we plot the different regions of origin as well as the age-gender distribution. We only considered participants who were fluent in English. We excluded participants taking any type of medication, tranquilizers, or psychotropic drugs (e.g., anti-depressants) as well as participants having any type of injury, disease, or long-term effects of a past disease (e.g., stiff neck, stroke, or tremor), which might affect body movements. All participants provided written informed consent before the start of the experiment and received monetary compensation.

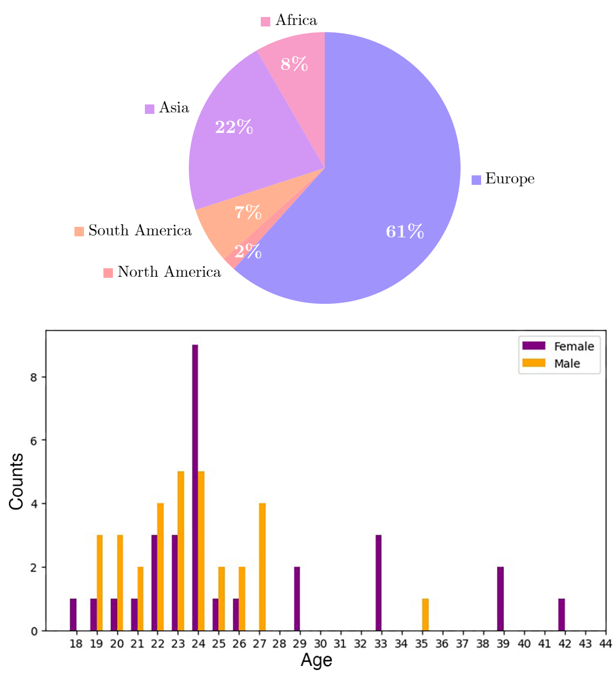


Figure 7: Distribution of the reported region of origin (top) and age/gender distribution (bottom) in COMOCAP.

4.1.2 Aparatus. Figure 8 illustrates the physical recording setup that we built for collecting COMOCAP. In this setup, two individuals were positioned to face each other. A screen was placed in front of them to show different text snippets and affective images. The participants were asked to read and act out the shown textual contents. They were further asked to have short conversations on the prompted images. We recorded RGB videos of the participants' faces at 60 FPS using an iPhone 13 Pro. Through the Live Link Face app [19], we further recorded ARKit blendshapes, which are commonly used in facial animation. Besides the face cameras, we also placed an iPhone stereo rig in front of the participants as a source for body motion (not used in this work). The audio was recorded at 48 kHz using two Rode Wireless Go II clip-on microphones. The data collection was approved by ETH Zurich's ethics committee and we will release parts of the dataset in the future.

Scripted Sentences. We randomly sampled 10 phonetically balanced lists from the Harvard Sentences [39]. Each list contains 10 sentences. Participants were asked to read aloud the sentences from two randomly selected lists (counterbalanced), which took approximately 10 minutes. Each sentence was shown on a screen in front of them. To ensure accurate pronunciation, participants listened to a speech file (synthesized using AudioStack [3]) before recording a sentence. As an example, Table 2 contains the sentences from the second Harvard list [39]. In total, we recorded 240,110 frames ($\approx$ 67 minutes) of active speech over 1,200 sequences (mean = 200 frames, SD = 73 frames). On average, each Harvard list is spoken by 12 participants.

Table 2: Second list of the Harvard Sentences Database [39]. Each list is phonetically balanced.

List	Harvard Sentences
2	The boy was there when the sun rose. A rod is used to catch pink salmon. The source of the huge river is the clear spring. Kick the ball straight and follow through. Help the woman get back to her feet. A pot of tea helps to pass the evening. Smoky fires lack flame and heat. The soft cushion broke the man's fall. The salt breeze came across from the sea. The girl at the booth sold fifty bonds.

Scripted Dialogs. Pairs of participants were acting out a subset of dialogs from DailyDialog [31] where each dialog turn is labeled in terms of emotion and intention. Since DailyDialog has an imbalanced label pair (i.e., emotion-intention) distribution, we developed a randomized downsampling algorithm that approximates uniformity across label pairs in the subset (see Section 4.1.3). We used this algorithm to sample 100 dialogs from DailyDialog. An example dialog can be found in Table 3. Each pair of participants acted out between 15 and 30 randomly selected dialogs (25 minutes on average). In total, we recorded 715,590 frames ($\approx$ 199 minutes) of active speech (mean = 237 frames, SD = 174 frames), and 7.59 participant pairs per dialog on average.

Free Dialogs. To capture non-acted, realistic performances with varying emotions, participants were asked to have short free conversations on affective images from the GAPED dataset [16] for 25 minutes. GAPED contains 730 images with ratings for valence, arousal, and the alignment of the depicted scene with internal (moral) and external (legal) norms. We performed k-means clustering with k = 3 (corresponding to a positive, neutral, and negative category) on the valence and arousal ratings of the images. From each of the three clusters, we randomly chose three samples that had a small standard deviation in terms of valence and arousal (i.e.,

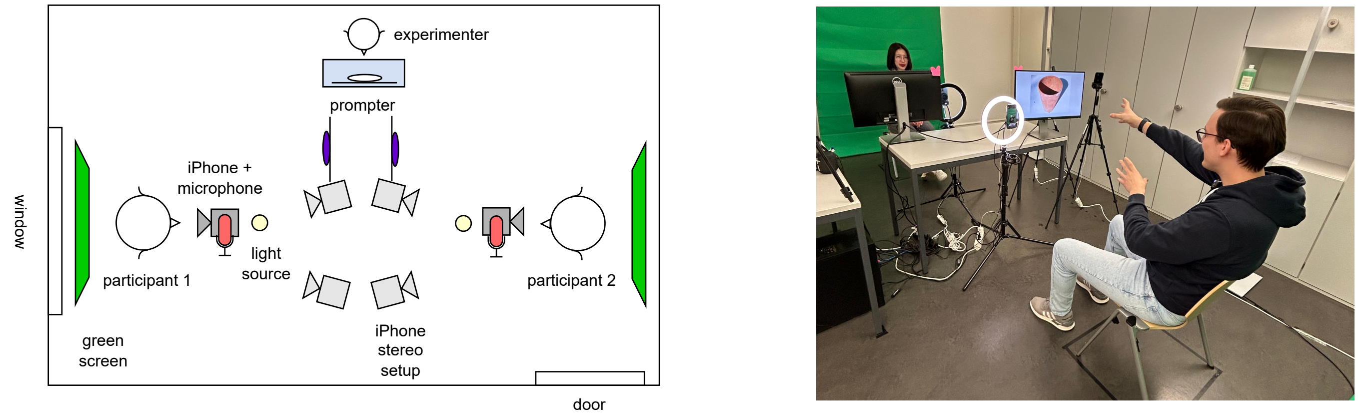


Figure 8: Left: Sketch of the recording setup. Right: Sample scene from the free dialogs (iii) where participants were shown a neutral image stimulus (i.e., a flower pot).

Table 3: Sample dialog from DailyDialog [31] including the ground truth dialog act and emotion labels.

ID	Utterance (DailyDialog)	Dialog Act	Emotion
A	Waiter!	Directive	No emotion
B	I'll be with you in a second. Uh... Yes, ma'am ?	Commissive	No emotion
A	This is not what I asked for. I'm afraid.	Inform	Sadness
B	Oh, I'm so sorry. May I ask what you ordered again?	Directive	Sadness
A	Yes. What I ordered is roast beef, not roast beef sandwiches.	Commissive	No emotion

Table 4: Sample conversation between two participants on the neutral image (i.e., a flower pot) from Figure 8 including the predicted dialog act labels.

ID	Utterance (Custom)	Predicted Dialog Act
A	Is that a flower pot or something?	Question
B	Yeah.	Inform
A	It seems so, because of the color and the dust.	Inform
B	Yeah, you're right.	Inform
...	...	...

9 images in total). The images were upscaled to 2560×1920 pixels using ESRGAN [50]. Every participant pair was presented with all 9 images in a counterbalanced order with respect to the category.

We display a sample scene from the free dialogs in Figure 8. Two participants discuss a neutral image stimulus (i.e., a flower pot) from the GAPED dataset. The recordings from the free dialogs were diarized using pyannote.audio [7, 8]. Google Speech-to-Text was used to transcribe the diarized snippets. Nevertheless, manual correction was required to fix faulty diarizations and transcriptions.

To obtain dialog act labels, we used our multimodal classifier trained on DailyDialog [31] and DailyTalk [30] and performed inference on the free dialogs. In Table 4, we show a sample dialog on the affective image in Figure 8 as well as the predicted labels. In total, we recorded 2,240,391 frames (≈ 622 minutes) of active speech (mean = 401 frames, SD = 560 frames).

4.1.3 Downsampling Algorithm. In the following, we explain the downsampling algorithm that we developed to reduce the size of DailyDialog for our data collection while achieving close uniformity on label pairs. This is not a trivial task since a dialog always contains multiple labels due to its multiple turns. Let n be the number of unique emotion-intention label pairs in the entire set R, i.e., $n = 4 \cdot 7 = 28$ for DailyDialog. The algorithm encodes a dialog i in the form of a vector $v_i \in \mathbb{N}^n$ where each entry $j \in [1, n]$ represents the number of occurrences of label pair j in dialog i. Let further be $\mathrm{var}(v_i)$ the variance across the elements of v_i. Our goal is to sample $S \subseteq R$ with $|S| << |R|$ such that

$$S = \arg\min_{v_i \in S} \mathrm{var}\left(\sum_{i=0}^{|S|-1} v_i\right). \tag{7}$$

We believe this problem is NP-complete (reduction to the knapsack problem) and thus only provide an approximate solution. Our algorithm randomly selects an initial subset S_0. In each iteration t, the algorithm draws a new sample $v_i' \in R \setminus S_{t-1}$ and replaces a randomly selected $v_i \in S_{t-1}$, i.e.,

$$S_t = S_{t-1} \setminus \{v_i\} \cup \{v_i'\}, \tag{8}$$

iff. v_i' reduces $\mathrm{var}(S_t)$. While the majority label pair *(no emotion, inform)* is represented approximately 2,600 times as often as the minority label *(disgust, directive)* in DailyDialog, we can reduce this factor to 38 in S. However, the majority and minority labels change to *(no emotion, commissive)* and *(disgust, commissive)*, respectively.

4.1.4 Procedure. Upon arrival, participants read the written information form. After giving their written informed consent, they performed the three types of recordings as described in Section 4.1.2. There were always two participants at the same time. Finally, they filled in an exit questionnaire and received their monetary compensation. This took approximately 90 minutes in total.

5 RESULTS

5.1 Quantitative Results

Table 5 shows the performance of our multimodal model on all three datasets. We used a modality fade-in approach during training. We trained on DailyDialog, evaluated the network's performance on its test set, then fine-tuned on DailyTalk, evaluated on its test set again, and finally fine-tuned on the scripted dialogs in COMOCAP and evaluated on its test set. Thus, our model always uses all modalities of the current dataset. As a baseline, we trained the unimodal state-of-the-art architecture developed by He et al. [25]. For a fair comparison, we also trained the baseline on DailyDialog first and fine-tuned it on DailyTalk and the scripted dialogs in COMOCAP afterwards. Here, we again freeze the text branch of the baseline and only train the decoding layers. Additionally, Table 5 shows the performance on DailyDialog for state-of-the-art *multi-task* approaches (DCR-Net [34], Co-GAT [35], and ChatGPT 3-shot, P.E [55]) as well as the best performing unimodal *single-task* approaches (CASA [37], and WEAKDAP [12]). Our model is superior to all other methods in the unimodal case (DD column). This indicates that the Transformer-based text encoder learns descriptive patterns that can be successfully decoded to dialog act predictions. The small improvement compared to the baseline [25] is most likely caused by our model's more powerful decoder. When fine-tuning on DailyTalk, our model successfully integrates acoustic features and outperforms the text-only model, leading to a relative improvement in the Macro F1 score of 1.24% compared to the baseline. This aligns with the findings from other bi-modal approaches [22, 33], which have been trained on different datasets comprising different dialog act labels. Since COMOCAP contains more speakers, further training the audio encoder adds more variance, which can lead to a more robust model. The results obtained on this dataset show an even greater relative improvement in performance (1.7%).

5.1.1 *Modality Ablation Study.*

To better understand the significance of each modality for DAC, we trained our model on subsets of the available modalities in COMOCAP (scripted dialogs) and report the corresponding performance in Table 6. For this ablation, we did not perform any pre-training on DailyDialog nor DailyTalk. For classifying dialog acts in an unimodal setting, text is the most discriminative modality, followed by audio. When pairing two modalities, the combination of text and audio leads to the highest Macro F1 score. Moreover, adding video to text does not have a significant effect, indicating that facial expressions may not represent an important cue for detecting intent. Interestingly, combining audio and video worsens the performance compared to the unimodal approaches. Lastly, the model that uses all three modalities achieves the best performance. While text and audio may be sufficient for detecting intent, visual inputs can stabilize predictions when the training set is small like the scripted dialogs in COMOCAP. Here, the labeled fraction that can be used contains only 100 unique dialog contents.

5.2 Qualitative Results

To gain a better understanding of how the additional modalities influence our model's decision-making, we examined individual predictions made by the multi- and unimodal (text-only) models. The scripted dialogs in COMOCAP contain several duplicated conversations uttered by distinct participants. Consequently, the audio and video recordings linked with these conversations are unique. Therefore, utilizing samples from this category allows for an exploration of how the behavior of the multimodal model evolves in response to the varying additional input modalities.

We manually analyzed the predictions on utterances for speakers where the performance of our multimodal model deviated most from the unimodal baseline. This includes cases where our multimodal model accurately predicts the ground-truth label, whereas the unimodal model fails, and vice versa. We discovered that most of those utterance are composed of multiple sentences with varying intent. Table 7 shows two examples where the DAC result changes with integrating the additional modalities. In the first utterance, the sentence "He is not breathing and there's no pulse." is related to the function of *inform*, while "Call 911." represents a *directive* message. In this case, including audio and video in addition to text helps our model to correctly predict the ground-truth label. For the second utterance, the first sentence "Yeah, it's boring." corresponds to an *inform* message, while "I'd rather read something more exciting." represents a *commissive* message. Here, the incorporation of additional modalities negatively impacted the predictive performance.

Figure 9 displays the Root-Mean-Square energy of the audio signals for the described utterances (see Table 7). In both examples, it can be observed that the model chooses the intent of the sentence that has the highest RMS energy for forming a prediction. By emphasizing the sentence "Call 911.", the speaker expresses the urgency of this message. Therefore, the intent of this sentence becomes crucial and can be considered as the principal function of the whole utterance. Furthermore, this means that the multimodal model is not necessarily wrong with respect to the dialog act of the second utterance. Instead, the faulty prediction may arise from

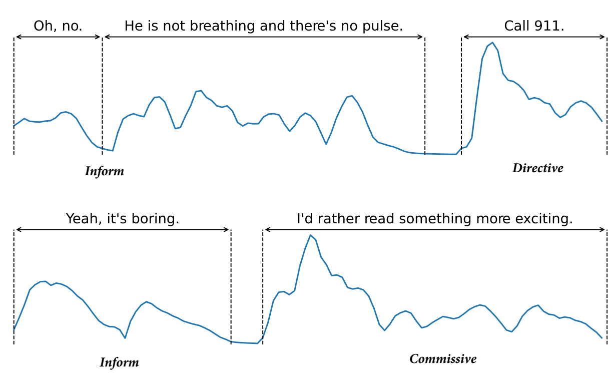


Figure 9: Acoustic energy of the two utterances in Table 7 in form of the Root-Mean-Square (RMS) energy of the audio signals. The intent of the sentence with the highest RMS energy corresponds to the multimodal model's prediction.

Table 5: Performance on the test sets of DailyDialog (DD), DailyTalk (DT), and COMOCAP (CM) when using modality fade-in training for the unimodal baseline [25] and our proposed multimodal network. We only used the scripted dialogs from COMOCAP. Precision, recall, and F1 score are computed using macro-averaging. Our model improves the baseline on all test sets in terms of F1 score and achieves better performance compared to other state-of-the-art methods.

	Accuracy (↑)			Precision (↑)			Recall (↑)			F1 (↑)		
	DD	DT	CM	DD	DT	CM	DD	DT	CM	DD	DT	CM
DCR-Net* [34]	-	-	-	79.10	-	-	79.00	-	-	79.10	-	-
Co-GAT* [35]	-	-	-	81.00	-	-	78.10	-	-	79.40	-	-
ChatGPT 3-shot, P.E† [55]	84.00	-	-	-	-	-	-	-	-	72.00	-	-
CASA* [37]	-	-	-	77.90	-	-	76.50	-	-	78.00	-	-
WEAKDAP† [12]	84.20	-	-	-	-	-	-	-	-	-	-	-
He et al. [25]	84.36	85.20	**80.59**	80.34	80.59	74.67	78.81	79.15	**82.63**	79.52	79.82	76.44
Ours	**84.50**	**85.79**	77.96	**80.44**	**81.72**	**79.13**	**78.96**	**80.02**	78.42	**79.66**	**80.81**	**77.74**

*Results reported from [35]

†Results reported from [55], P.E stands for prompt engineering

Table 6: Performance of our approach when ablating modalities in COMOCAP (scripted dialogs). Precision, recall, and F1 score are macro-averaged. In terms of accuracy and F1 score, we achieve the best performance utilizing all three modalities.

Modalities	Accuracy (↑)	Precision (↑)	Recall (↑)	F1 (↑)
Text	68.75	63.65	67.35	62.75
Audio	52.01	49.71	51.08	50.15
Video	35.53	37.45	36.37	35.00
Text + Audio	74.67	**76.45**	72.97	73.37
Text + Video	65.46	61.96	63.76	62.77
Audio + Video	36.18	34.32	29.68	30.21
Text + Audio + Video	**75.33**	75.68	**75.37**	**75.00**

Table 7: Predictions of the unimodal vs. multimodal model on test samples in COMOCAP (scripted dialogs) and the corresponding ground-truth (GT). We additionally propose fine-grained labels that may better reflect the varying intent in the input utterances.

Utterance	GT	Fine-grained	Unimodal	Multimodal
Oh, no. He is not breathing and there's no pulse. Call 911.	Directive	*Inform, Directive*	Inform	Directive
Yeah, it's boring. I'd rather read something more exciting.	Inform	*Inform, Commissive*	Inform	Commissive

a disagreement between the annotator's and voice actor's interpretation of the utterance. Hence, more fine-grained labels may be required for longer utterances.

5.3 Online Chatbot Conversations

Our work is intended for the development of enhanced conversational digital characters. In such scenarios, we posit that incorporating character state labels, like intent, that have been predicted for a chatbot's derived response can contribute to making synthesized speech [44] and animations more realistic. To provide rationale for integrating our DAC model into ECA dialog systems, we perform two additional analyses. First, we quantitatively report our model's performance on a modified version of the utilized datasets. These modifications are designed to mimic human-chatbot conversations, where we only have multiple modalities available on the user side. Secondly, since there may remain a domain gap to real human-chatbot conversations, we additionally provide qualitative results on real interactions with a conversational agent. In this context, we test the usability of our multimodal model for predicting dialog acts in real-time.

5.3.1 Quantitative Results. We adjusted DailyTalk and COMOCAP such that the two speakers are assigned a "user" and "chatbot" role. For utterances belonging to the user, we replaced the initial text with transcriptions generated from the audio recordings using the Whisper Speech-to-Text (STT) model [36] from OpenAI. In actual chatbot interactions where the user communicates with the chatbot through speech, there are no ground-truth transcriptions available. Instead, they need to be generated on the fly using an STT model. Through this type of dataset modification, we can analyze the impact of faulty speech transcriptions, which commonly occur in state-of-the-art STT methods used for human-character interactions. For utterances belonging to the chatbot, we keep the initial text but remove the corresponding audio and video recordings.

Table 8 shows the Macro F1 scores on different versions of the train (T) and evaluation (E) sets of DailyTalk and COMOCAP obtained by our multimodal model and the baseline by He et al. [25]. The superscript indicates that the respective set was modified to simulate human-chatbot interactions. For example, for " T / E* ", the model was trained on the original training dataset (full modalities for both interlocutors) and evaluated on the altered version (only text for one interlocutor, full modalities for the other).

The results show that the performance of our model and the baseline degrades when they are evaluated on the modified conversations. The fact that the Macro F1 scores obtained by the unimodal model also decrease suggests that the poor quality of automatically generated transcriptions negatively impacts DAC. Furthermore, the

Table 8: Macro F1 scores on modified versions (*) of the train (T) and evaluation (E) sets of DailyTalk and COMOCAP to simulate human-chatbot interactions. For the modified versions, we removed audio and video from the input of one interlocutor and replaced the text by STT transcriptions for the other interlocutor. The STT methods used in the pipeline negatively impact the performance on DAC. Nevertheless, our method still outperforms the baseline as long as the training and evaluation distributions align.

	Macro F1 (↑)					
Dataset	DailyTalk			COMOCAP		
	T / E	T / E*	T*/ E*	T / E	T / E*	T* / E*
He et al. [25]	79.82	**77.39**	77.57	76.44	**75.23**	75.67
Ours	**80.81**	77.18	**78.52**	**77.74**	74.33	**76.23**

performance of the unimodal model increases slightly when the training utterances are also altered (i.e., the model becomes more robust to transcription mistakes). This increase in performance is substantial in the case of the multimodal model. These results indicate that training on modified utterances is essential for robust DAC in human-character interactions. Furthermore, it shows that faulty transcriptions of state-of-the-art STT methods introduce non-negligible error sources in DAC. These are essential insights for developing robust dialog systems.

5.3.2 Online Inference. To mimic real online interactions with a conversational agent, we conversed with a GPT-3.5-based chatbot through speech, which was automatically transcribed and sent to the chatbot. The corresponding conversational user interface is shown in Figure 10. The transcription of a sample conversation and the dialog acts inferred on the chatbot's text and the user's text and

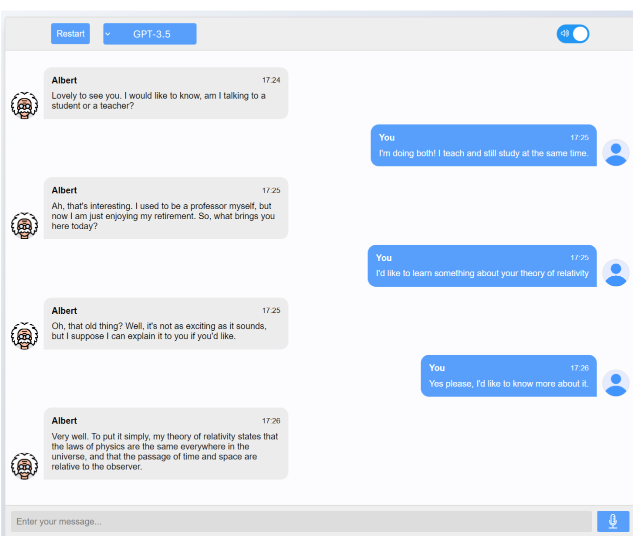


Figure 10: Conversational user interface for interactions with a GPT-3.5-based chatbot. Audio is recorded and transcribed to text, which is then sent to the chatbot.

Table 9: Online Inference on utterances between a human and a chatbot using our best pre-trained multimodal model. For the user's utterances, both text and audio inputs were used. For the chatbot, the model only uses text.

Speaker	Utterance	Predicted Dialog Act
Chatbot	Please make yourself at home. Ask me anything.	Directive
User	How do you like your coffee?	Question
Chatbot	With a bit of milk and sugar, it's a nice morning drink.	Inform
User	That's nasty! Sugar in coffee makes me vomit!	Inform
Chatbot	Ha! I understand. Milk and sugar can be difficult for some people.	Inform
User	Can you go and make me a black coffee?	Directive
Chatbot	Of course! Let me know if you like it.	Commissive
User	Oh, I didn't expect it to be this hot!	Inform
Chatbot	That's why I like to add a bit of cool water to my coffee.	Inform
User	You almost burned me. You should be more careful!	Directive
Chatbot	Yes, I will be more careful in the future.	Commissive

audio are shown in Table 9. The predicted dialog acts correspond to the expected intents of both the user's and the chatbot's utterances.

In Figure 11, we show the average time measurements of the previously described pipeline for DAC in human-chatbot interactions. This also includes video on the user side, which is not sent to the chatbot but only processed locally by our DAC model. Hence, it does not require any extensive transfer of sensitive data. For this, we consider an utterance of average length. Its audio signal is 4 seconds in length, the sentence contains 11 words, and 8 total frames are extracted from the video recording at equal intervals (2 per second). In total, less than 0.3 seconds are needed to obtain the intent of the utterance. Among these, 0.082 seconds are required to generate the transcription of the utterance and 0.037 seconds to get the timestamps for each word. Video pre-processing takes another 0.104 seconds. Finally, our multimodal model takes 0.04 seconds to make a prediction. Overall, this corresponds to 15% of the total inference time of 0.263 seconds. With these run-times, our approach can be used in online applications and is hence suitable for ECA dialog systems.

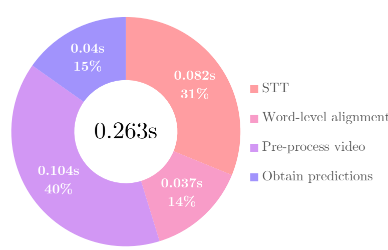


Figure 11: Time required for inference on a user's utterance (4 seconds, 11 words, 8 frames) incorporating all modalities in a user-chatbot interaction. The video pre-processing for our DAC model is the bottleneck of the pipeline. However, a prediction can still be made in real-time.

5.3.3 Privacy. When processing video data for DAC, privacy concerns may arise. However, our method runs locally and hence the video data is not transferred to external servers. Furthermore, our proposed method is tailored for dialog systems which inherently include video processing to enhance the interactive experience, e.g., allowing the character to react to user characteristics (see Figure 1 for an example). In such contexts, users might naturally anticipate the character's observational capabilities, potentially adjusting their privacy expectations accordingly. Given this, incorporating video data for DAC aligns seamlessly with the existing functionalities of these platforms and does not introduce additional complexity. Overall, however, we acknowledge a trade-off between the improvement in accuracy and privacy preservation.

5.4 Limitations

Our findings provide valuable insights into the impact of acoustic and visual features on DAC, supporting the hypothesis that a multimodal approach is superior for this task. However, there are certain limitations to the methods that we used in this work. First, we only had access to a comparably small number of labeled conversations comprising all three modalities. Therefore, further analysis on a larger dataset based on text, audio, and video may be required. Furthermore, apart from the scripted dialogs in COMOCAP, the used training datasets are highly imbalanced. Although this problem is partially solved by utilizing a weighted sampling technique, a balanced dataset would provide more diverse examples for the underrepresented classes. Moreover, the usefulness of visual cues is not fully exploited in our approach because of the way the video recordings are processed and how their features are extracted. Due to spatial and temporal constraints, we had to compromise visual information by reducing the videos to a small number of frames. In addition, due to the limitations of our recruiting pool, COMOCAP contains a bias in terms of age and region of origin. This may impact the type of conversations present in the free dialogs of COMOCAP as well as the distribution of auditory and visual features. Lastly, existing research has primarily focused on the role of dialog acts in facilitating response selection and enhancing speech synthesis. However, the impact of dialog acts on animation synthesis for ECAs remains unexplored. This aspect will be a focus of future research.

6 CONCLUSION

Our work represents an important step towards the development of enhanced conversational digital characters. We developed a multimodal network that uses text, audio, and video to predict dialog acts in conversations. Leveraging the cues offered by each communication channel, our model can accurately classify the utterances with respect to the functions they serve in the dialog (inform, question, directive, or commissive). We collected a novel multimodal dataset to train our network. We showed that our model outperforms previous unimodal networks by 1.7% Macro F1 score. Furthermore, we performed an ablation study where we demonstrated that using multiple input modalities reduces the number of required training samples due to faster learning. Lastly, we highlighted the robustness and versatility of our model by showing that it can predict dialog acts in human-chatbot conversations in real-time, taking less than 0.3 seconds on average. By having real-time access to the dialog acts associated with past utterances in the dialog, we lay the groundwork for research on enhanced speech and animation synthesis for digital characters. Our work can be easily integrated into existing dialog systems, providing a seamless enhancement to their functionality and capabilities.

ACKNOWLEDGMENTS

This work was supported by a Swiss National Science Foundation Grant under Grant No.: PZ00P2_216294.

REFERENCES

[1] Dilafruz Amanova, Volha Petukhova, and Dietrich Klakow. 2016. Creating Annotated Dialogue Resources: Cross-domain Dialogue Act Classification. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Portorož, Slovenia, 111–117. https://aclanthology.org/L16-1017

[2] Elisabeth André and Catherine Pelachaud. 2010. Interacting with embodied conversational agents. *Speech technology: Theory and applications* 1 (2010), 123–149.

[3] Ltd AudioStack Aflorithmic Labs. 2023. AudioStack - AI Audio Production. https://aflorithmic.ai [Accessed: (02.10.2023)].

[4] Elisabetta Bevacqua, Sathish Pammi, Sylwia Julia Hyniewska, Marc Schröder, and Catherine Pelachaud. 2010. *Multimodal backchannels for embodied conversational agents*. Vol. 6356 LNAI. Springer Verlag, Berlin. 194–200 pages. https://doi.org/10.1007/978-3-642-15892-6{_}21

[5] Timothy Bickmore and Justine Cassell. 2005. Social dialongue with embodied conversational agents. *Advances in natural multimodal dialogue systems* 1 (2005), 23–54.

[6] Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1307

[7] Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*. International Speech Communication Association, Brno, Czechia, 3111–3115. https://doi.org/10.21437/Interspeech.2021-560

[8] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote.Audio: Neural Building Blocks for Speaker Diarization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Barcelona, Spain, 7124–7128. https://doi.org/10.1109/ICASSP40776.2020.9052974

[9] Shuyi Cao, Lizhen Qu, and Leimin Tian. 2021. Causal Relationships Between Emotions and Dialog Acts. In *2021 9th International Conference on Affective Computing and Intelligent Interaction, ACII 2021*. IEEE Press, Nara, Japan, 1–8. https://doi.org/10.1109/ACII52823.2021.9597428

[10] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 520–527. https://doi.org/10.1145/302979.303150

[11] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. 2022. DistilHuBERT: Speech Representation Learning by Layer-wise Distillation of Hidden-unit BERT. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Singapore, Singapore, 7087–7091. https://doi.org/10.1109/ICASSP43922.2022.9747490

[12] Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022. Weakly Supervised Data Augmentation Through Prompting for Dialogue Understanding. arXiv:2210.14169 [cs.CL]

[13] Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue Act Recognition via CRF-Attentive Structured Network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 225–234. https://doi.org/10.1145/3209978.3209997

[14] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase

Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[15] Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. AAAI Press, New York, NY, USA, 7594–7601.

[16] Elise S. Dan-Glauser and Klaus R. Scherer. 2011. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior research methods* 43 (2011), 468–477.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, USA, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[18] Filippo Domaneschi, Marcello Passarelli, and Carlo Chiorri. 2017. Facial expressions and speech acts: experimental evidences on the role of the upper face as an illocutionary force indicating device in language comprehension. *Cognitive processing* 18 (2017), 285–306.

[19] Epic Games. 2023. Live Link Face. Mobile Application. https://dev.epicgames.com/community/learning/tutorials/lEYe/unreal-engine-facial-capture-with-live-link

[20] Spandana Gella, Aishwarya Padmakumar, Patrick Lange, and Dilek Hakkani-Tur. 2022. Dialog Acts for Task Driven Embodied Agents. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek (Eds.). Association for Computational Linguistics, Edinburgh, UK, 111–123. https://doi.org/10.18653/v1/2022.sigdial-1.13

[21] Barbara J Grosz. 1982. Discourse analysis. *Sublanguage: Studies of language in restricted semantic domains* 1 (1982), 138–174.

[22] Yue Gu, Xinyu Li, Shuhong Chen, Jianyu Zhang, and Ivan Marsic. 2017. Speech Intention Classification with Multimodal Deep Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10233 LNAI. Springer Verlag, New York, NY, USA, 260–271. https://doi.org/10.1007/978-3-319-57351-9{_}30

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, 770–778. https://doi.org/10.1109/CVPR.2016.90

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2016-December. IEEE Press, Las Vegas, 770–778. https://doi.org/10.1109/CVPR.2016.90

[25] Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. 2021. Speaker Turn Modeling for Dialogue Act Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Stroudsburg, PA, USA, 2150–2157. https://doi.org/10.18653/v1/2021.findings-emnlp.185

[26] Jill House. 2006. Constructing a context with intonation. *Journal of Pragmatics* 38, 10 (10 2006), 1542–1558. https://doi.org/10.1016/j.pragma.2005.07.005

[27] Daniel Jurafsky, Elizabeth Shriberg, Barbara Fox, and Traci Curl. 1998. Lexical, Prosodic, and Syntactic Cues for Dialog Acts. In *Discourse Relations and Discourse Markers*. The Association for Computational Linguistics, Montreal, Quebec, Canada. https://aclanthology.org/W98-0319

[28] Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with CRF. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) *(AAAI'18/IAAI'18/EAAI'18, Vol. 32)*. AAAI Press, New Orleans, 8 pages.

[29] Harshit Kumar, Arvind Agarwal, and Sachindra Joshi. 2019. A Practical Dialogue-Act-Driven Conversation Model for Multi-Turn Response Selection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 1980–1989. https://doi.org/10.18653/v1/D19-1205

[30] Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. DailyTalk: Spoken Dialogue Dataset for Conversational Text-to-Speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, IEEE, Rhodes Island, Greece, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095751

[31] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Greg Kondrak and Taro Watanabe (Eds.). Asian Federation of Natural Language Processing, Taipei, Taiwan, 986–995. https://aclanthology.org/I17-1099

[32] 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:1907.11692 [cs.CL]

[33] Daniel Ortega and Ngoc Thang Vu. 2018. Lexico-Acoustic Neural-Based Models for Dialog Act Classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB, Canada). IEEE Press, Calgary, AB, Canada, 6194–6198. https://doi.org/10.1109/ICASSP.2018.8461371

[34] Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020. DCR-Net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, Vol. 34. AAAI Press, Palo Alto, 8665–8672. https://doi.org/10.1609/aaai.v34i05.6391

[35] Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-GAT: A Co-Interactive Graph Attention Network for Joint Dialog Act Recognition and Sentiment Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. AAAI Press, Palo Alto, CA, USA, 13709–13717.

[36] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning* (, Honolulu, Hawaii, USA,) *(ICML'23, Vol. 202)*. JMLR.org, , Honolulu, Hawaii, USA, Article 1182, 27 pages.

[37] Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.), Vol. 1. Association for Computational Linguistics, Minneapolis, Minnesota, 3727–3733. https://doi.org/10.18653/v1/N19-1373

[38] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2017. The Influence of Context on Dialogue Act Recognition. arXiv:1506.00839 [cs.CL]

[39] E. H. Rothauser. 1969. IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics* 17, 3 (1969), 225–246.

[40] Tulika Saha, Dhawal Gupta, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Emotion Aided Dialogue Act Classification for Task-Independent Conversations in a Multi-modal Framework. *Cognitive Computation* 13, 2 (3 2021), 277–289. https://doi.org/10.1007/s12559-019-09704-5

[41] John R. Searle. 1969. *Speech Acts*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139173438

[42] Guokan Shang, Antoine Tixier, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2020. Speaker-change Aware CRF for Dialogue Act Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 450–464. https://doi.org/10.18653/v1/2020.coling-main.40

[43] Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. 2020. Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion. *IEEE Access* 8 (2020), 176274–176285. https://doi.org/10.1109/ACCESS.2020.3026823

[44] Vivek Kumar Rangarajan Sridhar, Ann K Syrdal, Alistair Conkie, and Srinivas Bangalore. 2011. Enriching Text-to-Speech Synthesis Using Automatic Dialog Act Tags.. In *Proc. Interspeech 2011*. International Speech Communication Association, Florence, Italy, 317–320. https://doi.org/10.21437/Interspeech.2011-119

[45] Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26, 3 (2000). https://doi.org/10.1162/089120100561737

[46] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26, 3 (2000), 339–373.

[47] Dinoj Surendran and Gina Anne Levow. 2006. Dialog Act Tagging With Support Vector Machines and Hidden Markov Models. In *Proc. Interspeech 2006*, Vol. 4. Citeseer, IEEE Press, Pittsburgh, PA, USA, paper 1831–Wed2FoP.12. https://doi.org/10.21437/interspeech.2006-535

[48] David R. Traum and Christine H. Nakatani. 1999. A Two-level Approach to Coding Dialogue for Discourse Structure: Activities of the 1998 DRI Working Group on Higher-level Structures. In *Proceedings of the ACL'99 Workshop Towards Standards and Tools for Discourse Tagging*. Association for Computational Linguistics, College Park, Maryland, USA, 101–108. https://aclanthology.org/W99-0313

[49] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1. IEEE, Kauai, HI, USA, I–I.

https://doi.org/10.1109/cvpr.2001.990517

[50] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*. Springer Cham, Munich, Germany.

[51] Yuqiao Wen, Guoqing Luo, and Lili Mou. 2022. An Empirical Study on the Overlapping Problem of Open-Domain Dialogue Datasets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 146–153. https://aclanthology.org/2022.lrec-1.16

[52] Bowen Xing and Ivor Tsang. 2022. DARER: Dual-task Temporal Relational Recurrent Reasoning Network for Joint Dialog Sentiment Classification and Act Recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3611–3621. https://doi.org/10.18653/v1/2022.findings-acl.286

[53] Kaicheng Yang, Hua Xu, and Kai Gao. 2020. CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis. In *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, Inc, Seattle, WA, USA, 521–528. https://doi.org/10.1145/3394171.3413690

[54] Jiahong Yuan, Mark Liberman, et al. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123, 5 (2008), 3878.

[55] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is ChatGPT Equipped with Emotional Dialogue Capabilities? arXiv:2304.09582 [cs.CL]