Neurocomputing 388 (2020) 102–109

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Dialogue systems with audio context

Tom Young[a], Vlad Pandelea[a], Soujanya Poria[b], Erik Cambria[a,*]

[a] Nanyang Technological University, Singapore
[b] Singapore University of Technology and Design, Singapore

ARTICLE INFO

Article history:
Received 1 October 2019
Revised 23 December 2019
Accepted 25 December 2019
Available online 14 January 2020

Communicated by Dr. Oneto Luca

Keywords:
Dialogue systems
Audio features
Language generation
Multimodality

ABSTRACT

Research on building dialogue systems that converse with humans naturally has recently attracted a lot of attention. Most work on this area assumes text-based conversation, where the user message is modeled as a sequence of words in a vocabulary. Real-world human conversation, in contrast, involves other modalities, such as voice, facial expression and body language, which can influence the conversation significantly in certain scenarios. In this work, we explore the impact of incorporating the audio features of the user message into generative dialogue systems. Specifically, we first design an auxiliary response retrieval task for audio representation learning. Then, we use word-level modality fusion to incorporate the audio features as additional context to our main generative model. Experiments show that our audio-augmented model outperforms the audio-free counterpart on perplexity, response diversity and human evaluation.

1. Introduction

In recent years, data-driven approaches to building conversation models have been made possible by the proliferation of social media conversation data and the increase of computing power. Based on a large amount of conversation data, very natural-sounding dialogue systems can be built by learning a mapping from textual context to response using powerful machine learning models [1–4]. Specifically in the popular sequence-to-sequence (Seq2Seq) learning framework, the textual context, modeled as a sequence of words from a vocabulary, is encoded into a context vector by a recurrent neural network (RNN). This context vector serves as the initial state of another RNN, which decodes the whole response one token at a time.

This setting, however, is oversimplified compared to real-world human conversation, which is naturally a multimodal process [5,6]. Information can be communicated through voice [7], body language [8] and facial expression [9]. In some cases, the same words can carry very different meanings depending on information expressed through other modalities.

In this work, we are interested in audio signals in conversation. Audio signals naturally carry emotional information. For example, "Oh, my god!" generally expresses surprise. Depending on the voice shade, however, a wide range of different emotions can also be carried, including fear, anger and happiness. Audio signals can have strong semantic functions as well. They may augment or alter the meaning expressed in text. For example, "Oh, that's great!" usually shows positive attitude. With a particular voice shade of contempt, however, the same utterance can be construed as sarcastic. Stress also plays a role in semantics: "I think *she* stole your money" emphasizes the speaker's opinion on the identity of the thief while "I think she stole *your* money" emphasizes the speaker's opinion on the identity of the victim.

Therefore, while identical from a written point of view, utterances may acquire different meanings based solely on audio information. Empowering a dialogue system with such information is necessary to interpret an utterance correctly and generate an appropriate response.

In this work, we explore dialogue generation augmented by audio context under the commonly-used Seq2Seq framework. Firstly, because of the noisiness of the audio signal and the high dimensionality of raw audio features, we design an auxiliary response classification task to learn suitable audio representation for our dialogue generation objective. Secondly, we use word-level modality fusion for integrating audio features into the Seq2Seq framework. We design experiments to test how well our model can generate appropriate responses corresponding to the emotion and emphasis expressed in the audio.

In summary, this paper makes the following contributions:

(i) To the best of our knowledge, this work is the first attempt to use audio features of the user message in neural conversation generation. Our model outperforms the baseline audio-free model in terms of perplexity, diversity and human evaluation.

* Corresponding author
E-mail address: cambria@ntu.edu.sg (E. Cambria).

https://doi.org/10.1016/j.neucom.2019.12.126
0925-2312/

(ii) We perform extensive experiments on the trained model which show that our model captures the following phenomena in conversation: (1) Vocally emphasized words in an utterance are relatively important to response generation. (2) Emotion expressed in the audio of an utterance has influence on the response.

2. Related work

Massive text-based conversation data has driven a strong interest in building dialogue systems with data-driven methods. The Seq2Seq model, in particular, has been widely used due to its success in text generation tasks such as machine translation [10], video captioning [11] and abstractive text summarization [12]. Seq2Seq employs an encoder–decoder framework, where the conversational context is encoded into a vector representation and, then, fed to the decoder to generate the response [13]. Under the text-based Seq2Seq framework, a large number of works have been done on improving the content quality of the response, such as diversity promotion [2,14], integrating emotional information [15], and handling unknown words [16].

In contrast to the popular text-only assumption, however, human conversation naturally involves multiple modalities.

Firstly, the context of conversation can be multimodal. For example, in image-grounded conversation [17], two interlocutors generate conversations based on a shared image. For this task, visual features of the image need to be infused into the context vector. Alamri et al. [18] proposed Visual Scene-aware Dialogs, a scenario where the dialogue system discusses dynamic scenes with humans. A scene, in the form of a short video, is presented to the interlocutors as the conversational context. For this task, Hori et al. [19] incorporated techniques for multimodal attention-based video description into an end-to-end dialogue system. Audio and visual features that come from deep video description models are used to augment the context vector. Saha et al. [20] proposed a large domain-aware multimodal conversation dataset where shoppers and sales agents converse about products in the fashion domain. Each conversational turn is composed of text and corresponding images being referred to. For this scenario, Agarwal et al. [21] proposed a multimodal extension to the Hierarchical Recurrent Encoder–Decoder (HRED) [22] for in-turn multimodality and multi-turn context representation.

Secondly, human conversation itself involves multiple channels of information. Voice, body language and facial expressions all play roles in conversation. In an ideal human-machine conversational system, machines should understand this multimodal language. In the literature, this information has seen use in conversation analysis. Yu [23] proposed to model user engagement and attention in real time by leveraging multimodal human behaviors, such as smiles and speech volume. Gu et al. [6] performed emotion recognition, sentiment analysis, and speaker trait analysis on conversation data using a hierarchical encoder that formulates word-level features from video, audio, and text data into conversation-level features with modality attention.

Our method of word-level modality fusion has already seen use in multimodal sentiment analysis. In [24], the RNN, which acts as the utterance encoder, takes a concatenation of audio, video and text features as input at every time step. On the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [24,25] showed considerable improvement on dialogue emotion classification accuracy by integrating audio features. This result motivates our work – since incorporating audio features improves emotion classification accuracy in conversation and emotion is important to response generation [15], we hypothesize that incorporating audio features improves response generation.

3. Model

3.1. Audio representation learning

Raw features extracted from audio sequences are high-dimensional and noisy. They are not suited as direct input to the dialogue generative model. Therefore, we need an audio representation learning method to reduce audio feature dimensions and also make it suitable for the dialogue generation task.

For this purpose, we design an auxiliary response classification task based solely on audio features.

Specifically, we construct a set of $<$ *context, response, label* $>$ triples, where *label* is binary indicating whether the context and response combination comes from a real conversation dataset D or is randomly assembled. The goal of this task is to predict *label* based on the $<$ *context, response* $>$ pair.

Following [26], our classification model is defined as:

$$f(c, r) = sigmoid(\boldsymbol{c}^T \boldsymbol{W} \boldsymbol{r}), \tag{1}$$

where $\boldsymbol{c}$ and $\boldsymbol{r}$ are representations of the context c and response r respectively. Matrix $\boldsymbol{W}$ is model parameter.

We use a universal sentence encoder [27] for the representation of response $\boldsymbol{r}$. For the purpose of finding the best audio context representation, $\boldsymbol{c}$ is determined only by audio features $\boldsymbol{a}_i$ of individual words in the context:

$$\boldsymbol{c} = avg(P(\boldsymbol{a}_i)), i \in [0, L), \tag{2}$$

where P is a perceptron and L is the number of words in the context. The model is shown in Fig. 1.

This model is trained on a conversation dataset D for best classification accuracy using mean squared loss between *label* and $f(c, r)$ in Eq. (1). After training, the output of the perceptron $\tilde{\boldsymbol{a}}_i = P(\boldsymbol{a}_i)$ is taken as the word-level audio representation used in the generative dialogue systems.

3.2. Audio-augmented Seq2Seq model

We build upon the general encoder–decoder framework which is based on sequence-to-sequence (Seq2Seq) learning [28]. The encoder represents a user message (context) $X = x_1 x_2 \dots x_n$ with hidden representations $\boldsymbol{H} = \boldsymbol{h}_1 \boldsymbol{h}_2 \dots \boldsymbol{h}_n$, which is briefly defined as below:

$$\boldsymbol{h}_n = \mathbf{LSTM_E}(\boldsymbol{h}_{n-1}, \boldsymbol{e}(x_n)), \tag{3}$$

where $\mathbf{LSTM_E}$ is a long short-term memory (LSTM) network [29]. $\mathbf{E}$ denotes encoder. The decoder takes as input a context vector $\tilde{\boldsymbol{c}}_{t-1}$ produced by an attention mechanism and the embedding of a previously decoded word $\boldsymbol{e}(y_{t-1})$, and updates its state $\boldsymbol{s}_t$ using another LSTM:

$$\boldsymbol{s}_t = \mathbf{LSTM_D}(\boldsymbol{s}_{t-1}, [\tilde{\boldsymbol{c}}_{t-1}; \boldsymbol{e}(y_{t-1})]), \tag{4}$$

where $\mathbf{D}$ denotes decoder. The decoder generates a token by sampling from the output probability distribution which is determined by $\tilde{\boldsymbol{c}}_t$.

Following [24], we use a simple word-level embedding concatenation method for integrating audio features into word representation:

$$\boldsymbol{e}(x_n) = [\boldsymbol{w}_n; \tilde{\boldsymbol{a}}_n], \tag{5}$$

where $\boldsymbol{w}_n$ is the traditional word embedding and $\tilde{\boldsymbol{a}}_n$ is the word-level audio representation. Thus, in our new Audio-Seq2Seq model (Fig. 2), the word representation contains both textual and audio information.

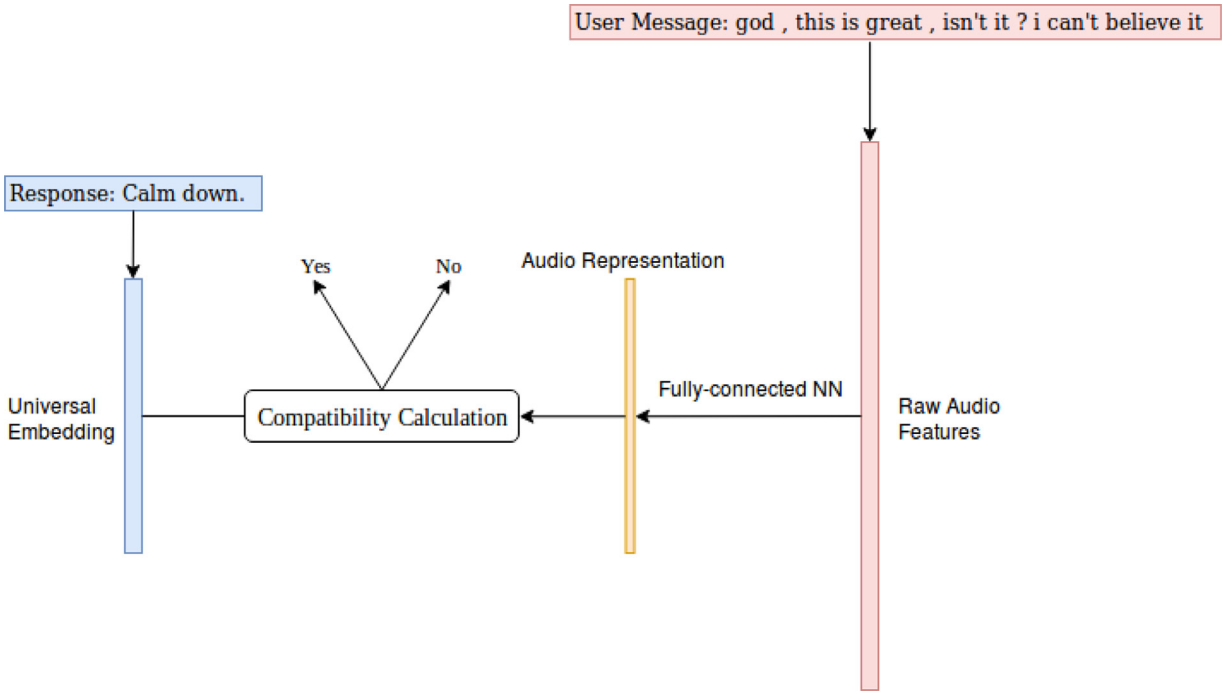


Fig. 1. A response classification model is used as the auxiliary task for audio representation learning.

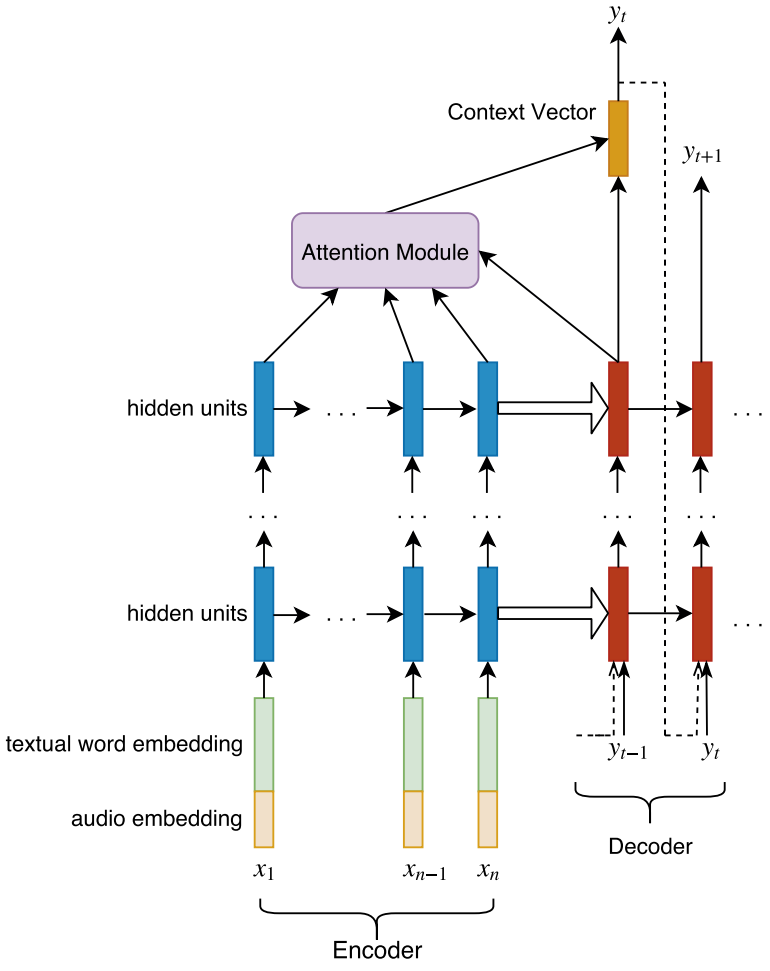


Fig. 2. Audio-Seq2Seq model.

4. Experiments

4.1. Dataset

Most of the existing and consolidated datasets used in dialogue system related research come with textual content only [26,30]. The predominance of text-only datasets can be seen as a consequence of both the ease with which this type of data can be acquired, and the lack of a real demand of multimodal conversation data until recent times. Fortunately, along with the growing interest in multimodal systems, there has also been an increase in the proliferation of datasets fit for our task. We experiment with two such datasets, the IEMOCAP [25] and the Multimodal EmotionLines Dataset (MELD) [31].

IEMOCAP was designed with the main intent of providing a corpus of dyadic conversations capable of conveying emotions. Two types of dialogue sessions were created for IEMOCAP to achieve this task: scripted and spontaneous sessions. In the scripted case, two actors, a male and a female, were asked to rehearse some previously memorized scripts, as this supposedly leads to a more genuine expression of emotions than directly reading off a script. In the spontaneous case, the actors were given more liberty to use their own words to discuss about selected emotion-evoking topics. This supposedly allows the actors to express more natural emotions. The dataset contains a total of 10,039 utterances with their corresponding audio segments.

MELD is a dataset containing utterances from the TV series *Friends*. For each utterance multimodal information in the form of text, audio and video is provided. MELD consists of 1433 dialogues for a total of 13,708 utterances.

These two datasets, IEMOCAP in particular, are suitable for our purposes as the audio component is strongly representative of the speaker's emotional state and plays a pivotal role in the meaning to be conveyed.

4.2. Experiment details

4.2.1. Data preprocessing

From IEMOCAP and MELD set of dialogues we extract $<$*sentence, response*$>$ pairs by taking successive utterances within individual dialogues. Formally, from dialogue $d_i = \{u_1, \ldots, u_{n_i}\}$, where $u_1, \ldots, u_{n_i}$ are the utterances composing the dialogue, we extract the set of pairs $\{<u_1, u_2>, \ldots, <u_{n_i-1}, u_{n_i}>\}$. From the resulting pairs we create a vocabulary, for each of the datasets, containing only the terms with more than one occurrence in the respective corpus and that are present in the standard English vocabulary [32] and those that are not present in the English vocabulary but occur ten or more times in the dataset. Our final vocabulary sizes are 2171 for IEMOCAP and 3123 for MELD.

Further, we also remove all the sentences that denote the end of a dialogue. After these procedures we end up with a total of 7901 utterances for IEMOCAP and $12,274$ for MELD.

The audio segments provided within the datasets are given at a sentence granularity. We conduct word alignment and obtain word-level audio features with the following procedure:

(i) We first use the GENTLE forced aligner [33] to find the start and end timestamps of each word within a sentence.
(ii) Then, with OpenSMILE [34], we extract 6373 raw audio features for each word. We use the *IS13_ComParE.conf* configuration [35] that has been widely used in emotion recognition tasks [31,36], rendering it a suitable choice for our case, as emotion and response generation are closely tied concepts that influence one another in human conversation.

The shorter, faster-paced and overall more noisy dialogues of MELD result in the failure of GENTLE to correctly align the words, or in the inability of OpenSMILE to extract the audio features, of around 23% of the words, whereas for IEMOCAP only around 7% of the words are left without corresponding audio features. For these words we use zero vectors as features. We randomly sample utterances from the datasets to split into training and development sets. In Table 1, we report some of the most prominent statistics regarding the datasets we operate on.

Table 1
Initial number of utterances, number of utterances after preprocessing, average length of utterances, development set sizes, test set sizes and vocabulary sizes for IEMOCAP and MELD datasets.

	IEMOCAP	*MELD*
No. utt.	10,039	13,708
Preproc. no. utt.	7901	12,274
Avg utt. length	15.26	10.69
Dev. set size	1000	1174
Test set size	901	1000
Vocabulary size	2171	3123

4.2.2. Model training details

In our audio representation learning model (Section 3.1), the response sentence embedding given by the universal sentence encoder has size 4096. During the training process, the best audio representation extractor is obtained at the point when the classification accuracy on the development dataset is the highest.

We use the Seq2Seq model with Luong attention mechanism [37] as the backbone of our main audio-augmented model. It is a pruned version of the main model (Fig. 2) that does not use audio features in its word-level representation. After being trained on a large text-based conversation dataset, the resulting model parameters are transferred to the main model as initialization of its parameters corresponding to textual input.

A 3.3M Reddit Conversation Dataset [38] is used for this purpose. We filter it using the vocabularies previously created for the audio conversation datasets. Specifically a conversation pair $<u_j, u_{j+1}>$ is removed if it contains more than one out-of-vocabulary term.

Our Audio-Seq2Seq model uses textual word embeddings of size 100 while the audio representation has size 25, which finds justification both intuitively, as although audio plays a part in human conversation, text is still most important as it carries semantic information directly, and experimentally on the auxiliary response classification task (Section 4.3.1).

We follow [39] for most of the hyperparameter settings. All generation models are trained for 50,000 steps with batch size 256 after initialization with the pretrained model. The learning rate is set to 0.1. Dropout rate is 0.3. We use 2 layers of hidden units. Keeping audio representation fixed at 25, we search for the optimal text embedding dimension in 10, 25, 50, 100. 100 yields the best results. By manually inspecting the generated responses at different steps, we find that they are most natural-sounding when the models slightly overfit. In contrast, the models generate overly simple responses when the development perplexity is lowest. This is due to the fact that the audio conversation datasets are relatively small. We manually choose the best checkpoint for testing based on human perception of response quality on the development set after the models start overfitting.

4.3. Experiment results

4.3.1. Results on audio representation learning

The results are shown in Table 2. The fact that the accuracies are much higher than 50% indicates that audio features indeed carry information that is relevant to conversation. The accuracies show only a slight improvement in spite of a substantial increase

Table 2
Auxiliary response classification task accuracy varying the dimension of the audio representation.

	Dimension		
Dataset	*25*	*50*	*100*
IEMOCAP	59.4%	62.4%	61.8%
MELD	54.8%	54.8%	54.6%

Table 3
Statistics on IEMOCAP.

	Metric		
Model	*Perplexity*	*Diversity*	*Human preference*
Seq2Seq	36.83 ± 0.34	805 ± 10.5	44.4%
Audio-Seq2Seq	31.13 ± 0.31	831 ± 12.8	55.6%

Table 4
Statistics on MELD.

	Metric		
Model	*Perplexity*	*Diversity*	*Human preference*
Seq2Seq	47.83 ± 0.44	567 ± 8.7	46.5%
Audio-Seq2Seq	46.19 ± 0.49	629 ± 10.0	53.5%

in dimension moving up from 25, which is another reason why 25 is the size of the audio representation that is adopted in all the experiments.

4.3.2. Perplexity, diversity and human evaluation

Fully automated evaluation of non-task-oriented open-domain dialogue systems remains an open challenge [40]. Human judgment of response quality is still the most reliable criterion.

In this work we consider two automatic metrics in addition to human judgment. The *perplexity* of the model on a ground-truth response S is defined as:

$$per(S) = P(w_1, \ldots, w_n)^{(-\frac{1}{n})} = \sqrt[n]{\frac{1}{P(w_1, \ldots, w_n)}} = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(w_i|w_1, \ldots, w_{i-1})}}. \quad (6)$$

The *perplexity* of the model on a set of responses is the average over all ground-truth responses composing the set.

Diversity is defined as the number of unique words generated by the model over the test set. Lack of diversity and tendency to generate similar, short responses regardless of the different inputs is a notorious problem in generative conversational models [41]. A model that generates interesting and information-rich responses is characterized by high diversity. Automatic and human evaluation results are shown in Tables 3 and 4. ± value indicates standard deviation. We see that the Audio-Seq2Seq model achieves lower perplexity and higher diversity on both datasets.

A number of samples were manually selected from the test set for human evaluation. During this process, we only chose samples with high-quality user messages, considering both audio and text. A sample is excluded if the audio is noisy or the user message is too short or generic.

We conducted pair-wise comparisons between responses generated by the Audio-Seq2Seq model and the Seq2Seq model. Seven judges were recruited on the Amazon Mechanical Turk (AMT) platform. After listening to the audio of the user message, the judges were asked to select the best response from those generated by the two models. We asked the judges to follow two rules: (1) Prefer the response that is compatible with the tone of the user message. (2) Prefer longer and more informative responses. The results of the judges' evaluation are shown in Table 5.

Table 5
Number of responses by the corresponding model that are deemed as better according to AMT judges.

	Audio-Seq2Seq	*Seq2Seq*
IEMOCAP	354	283
MELD	359	313

Table 6
Samples produced by Audio-Seq2Seq. Considering the tone of the context, the responses Audio-Seq2Seq produced are more appropriate.

Context	*Seq2Seq Response*	*Audio-Seq2Seq Response*
because i know you're at a desk right now in your corporate building and you're going to send me a check for sixteen dollars and forty two cents !	right, ok.	oh, no, no, no, no, no. i don't know. i don't know. it's not what i can do.
what time is it ? god, this is great, isn't it ? i can't believe it.	yeah, it is	okay, all right. just just calm down
no. it's his newspapers is what it is now every month there's a new boy comes home and larry's going to be the next one.	what ?	alright. alright. alright just listen.

Table 6 shows successful cases where the responses produced by the Audio-Seq2Seq model suit the audio context better.[1] For example, in the second case, the voice of the user message shows excitement and restlessness, which is captured by the "calm down" in the response.

4.3.3. Emotional appropriateness

Audio features can be strongly indicative of the speaker's emotion, and thus have influence on the response. In order to quantitatively evaluate the ability of our model to generate emotionally appropriate responses, we design an experiment with an artificially constructed set of 200 audio message samples of two different emotional states. Specifically, 100 textual message samples are selected from the test set and for each sample we manufacture two audio segments of different emotions by availing ourselves of the MARY Text-To-Speech system (MaryTTS) [42].

Following Russell's Circumplex model of affect [43] we vary the valence dimensions of the synthesized audio segments. With arousal and valence in the range [0,1], we use a fixed arousal value of 0.9 combined with the two valence values 0.1 and 0.9. When valence = 0.9, the synthesized speech is fast and highly-pitched, exhibiting an excited emotional state. Whereas when valence = 0.1, the synthesized speech is slow and calm. Our Audio-Seq2Seq model generates two responses corresponding to those two audio segments of different emotion states. To evaluate how well a response matches an emotional state, we shuffle the two responses and ask human judges to match audio segments with the responses to see if the results agree with the model's.

This association task performed by three judges shows that human evaluation tends to agree with the responses generated by the model. Details are given in Table 7. Table 8 shows cases where the model seems to be able to perceive the emotional state of the speaker and adapt its response accordingly. When the audio expresses an excited state (valence = 0.9), the model is able to tune its response in a suitable manner. For instance in the first sample

[1] All utterance samples displayed in this paper have corresponding audio files in the supplementary material.

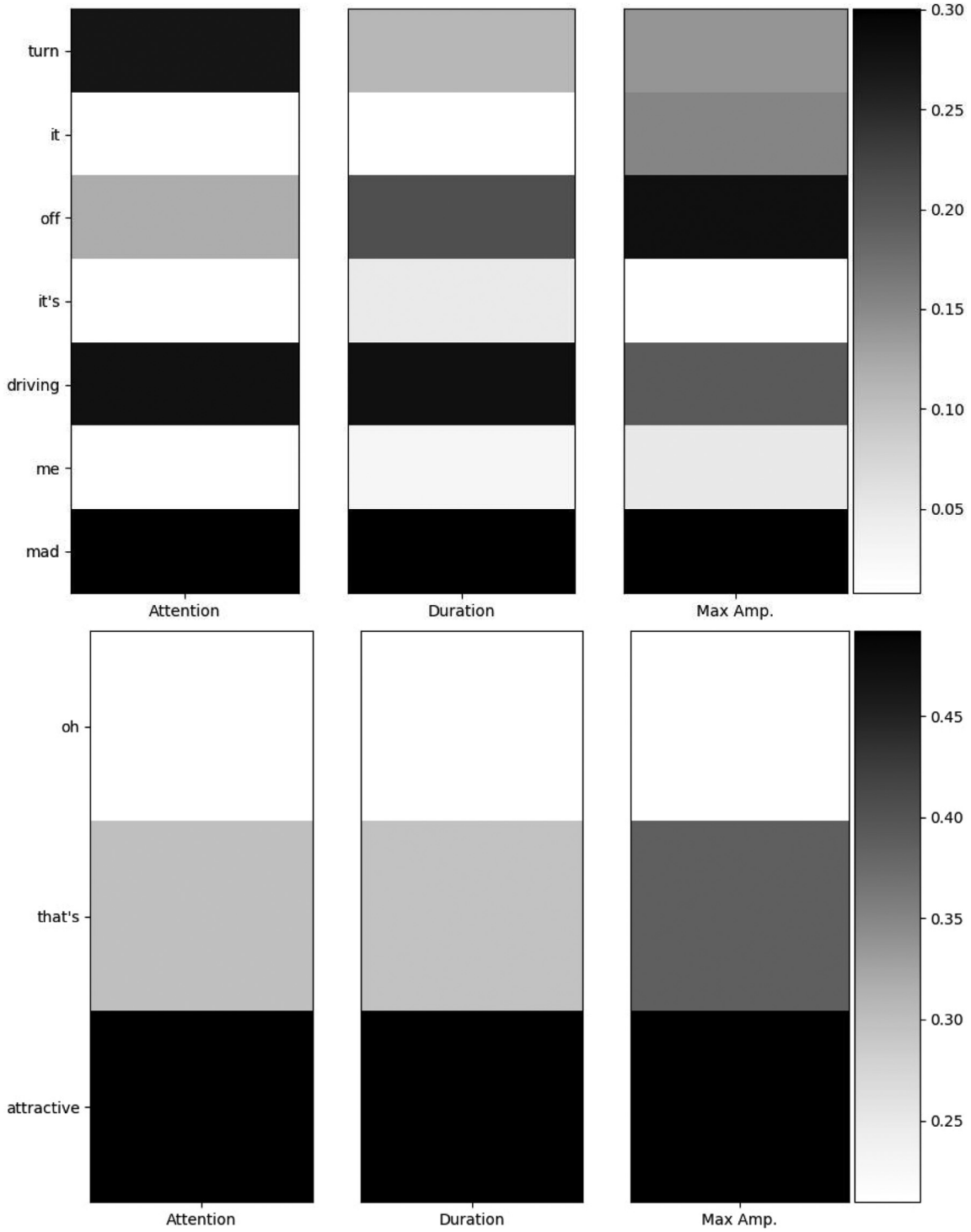


Fig. 3. The attention of a word in the source sequence is positively correlated with both duration and maximum amplitude.

Table 7
Percentage of cases on which the judges' verdicts agree or disagree with the model. We also report the cases for which the judges were not able to make an association.

Model	*Agree*	*Disagree*	*Cannot determine*
IEMOCAP	25.4%	15.1%	59.5%
MELD	28.2%	16.7%	55.1%

Table 8
MaryTTS samples.

Context	*Valence = 0.1*	*Valence = 0.9*
Turn it off it's driving me mad.	I won't.	Well, do try to control yourself darling.
okay that's helpful. thanks.	i've been trying to work this backwards	this is all this is unfair.

the second response shows a strong correlation with the excited and agitated state of the speaker by asking him to calm down. In the second sample, the higher rate with which the valence = 0.9 context is uttered due to the excited state makes the speaker sound less sincere thus eliciting a stuttered and complaining response as compared to the more composed and calm one when valence = 0.1.

Table 9
The correlation between word attention and duration/maximum amplitude on IEMOCAP.

IEMOCAP	*Pearson's r*	*Spearman's ρ*
Attention/duration	0.418	0.384
Attention/max amp.	0.096	0.128

Table 10
The correlation between word attention and duration/maximum amplitude on MELD.

MELD	*Pearson's r*	*Spearman's ρ*
Attention/duration	0.312	0.334
Attention/max amp.	0.094	0.069

4.3.4. Attention on vocally emphasized words

In a conversation, vocally emphasized words in an utterance are most important to information communication. To evaluate how well our model captures this phenomenon, we calculate the correlation between the volume/duration of the audio segments of words in the user message and the attention the words get during the generation process.

We take the length of the audio segment of an individual word as the *duration* of that word and *Maximum amplitude* is used to indicate *volume*.

For calculating attention on a word in the message, we sum all attention scores it gets during the response generation process. Specifically, for the generation of response word y_t, the attention score on message word x_i is a_{it}. For the generated response $[y_1, y_2 \dots y_n]$, the total attention on x_i is $a_i = \sum_{t=1}^{n} a_{it}$.

We normalize attention, duration and maximum amplitude by dividing them by average values over the message. Pearson and Spearman correlations are calculated on attention-duration and attention-maximum amplitude pairs. The results are shown in Tables 9 and 10. On both datasets our experiment shows relatively strong positive correlation between attention and duration. For attention and maximum amplitude, however, our calculation only shows slightly positive correlation. This implies that in our dataset, length is more indicative of a wordâs importance to the dialogue system than volume. However, it cannot be generalized without more experiments on more datasets.

Two examples are shown in Fig. 3. In the message “turn it off it’s driving me mad”, “off”, “driving” and “mad” are vocally emphasized. Accordingly, attention scores on those three words are relatively high. In a shorter example, “oh that’s attractive”, the word “attractive” contains the most semantic information. It is vocally emphasized and gets the most attention.

5. Conclusion

In this work, we augmented the common Seq2Seq dialogue model with audio features and showed that the resulting model outperforms the audio-free baseline on several evaluation metrics. It also captures interesting audio-related conversation phenomena.

Although only using text in dialogue systems is a good-enough approximation in a lot of scenarios, other modalities (i.e., video and audio) have to be integrated before automatic dialogue systems can reach human performance. Our work belongs to such a line of research that strives to build multimodal dialogue systems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Tom Young: Conceptualization, Methodology, Software, Writing - original draft. **Vlad Pandelea:** Data curation, Software. **Soujanya Poria:** Conceptualization. **Erik Cambria:** Supervision.

Acknowledgments

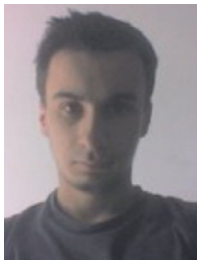
This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project #A18A2b0046).

References

[1] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, M. Huang, Augmenting end-to-end dialogue systems with commonsense knowledge, in: Proceedings of the 2018 AAAI, 2018, pp. 4970–4977.
[2] L. Shao, S. Gouws, D. Britz, A. Goldie, B. Strope, R. Kurzweil, Generating long and diverse responses with neural conversation models, CoRR (2017). abs/1701.03185.
[3] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, DialogueRNN: an attentive RNN for emotion detection in conversations, in: Proceedings of the 2019 AAAI, 2019, pp. 6818–6825.
[4] H. Xu, H. Peng, H. Xie, E. Cambria, L. Zhou, W. Zheng, End-to-end latent–variable task-oriented dialogue system with exact log-likelihood optimization, World Wide Web (2020) In press.
[5] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270.
[6] Y. Gu, X. Li, K. Huang, S. Fu, K. Yang, S. Chen, M. Zhou, I. Marsic, Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder, in: Proceedings of the 2018 ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 537–545.
[7] R. Kingdon, The semantic functions of stress and tone, ELT J. 3 (7) (1949) 178.
[8] J. Streeck, C. Goodwin, C. LeBaron, Embodied Interaction: Language and Body in the Material World, Cambridge University Press, 2011.
[9] A. Takeuchi, K. Nagao, Communicative facial displays as a new conversational modality, in: Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems, ACM, 1993, pp. 187–193.
[10] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, CoRR (2014). abs/1409.0473.
[11] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko, Sequence to sequence-video to text, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4534–4542.
[12] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, et al., Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond, arXiv:1602.06023 (2016).
[13] O. Vinyals, Q. Le, A Neural Conversational Model, arXiv:1506.05869 (2015).
[14] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan, A Diversity-Promoting Objective Function for Neural Conversation Models, arXiv:1510.03055 (2015).
[15] H. Zhou, M. Huang, T. Zhang, X. Zhu, B. Liu, Emotional chatting machine: emotional conversation generation with internal and external memory, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[16] J. Gu, Z. Lu, H. Li, V.O. Li, Incorporating copying mechanism in sequence-to-sequence learning, in: Proceedings of the 54th Annual Meeting of the Proceedings of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 2016, pp. 1631–1640.
[17] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G.P. Spithourakis, L. Vanderwende, Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation, arXiv:1701.08251 (2017).
[18] H. Alamri, V. Cartillier, R.G. Lopes, A. Das, J. Wang, I. Essa, D. Batra, D. Parikh, A. Cherian, T.K. Marks, et al., Audio Visual Scene-Aware Dialog (avsd) Challenge at dstc7, arXiv:1806.00525 (2018).
[19] C. Hori, H. Alamri, J. Wang, G. Winchern, T. Hori, A. Cherian, T.K. Marks, V. Cartillier, R.G. Lopes, A. Das, et al., End-to-End Audio Visual Scene-Aware Dialog Using Multimodal Attention-Based Video Features, arXiv:1806.08409 (2018).
[20] A. Saha, M.M. Khapra, K. Sankaranarayanan, Towards building large scale multimodal domain-aware conversation systems, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[21] S. Agarwal, O. Dusek, I. Konstas, V. Rieser, Improving Context Modelling in Multimodal Dialogue Generation, arXiv:1810.11955 (2018).
[22] I.V. Serban, A. Sordoni, Y. Bengio, A. Courville, J. Pineau, Building end-to-end dialogue systems using generative hierarchical neural network models, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[23] Z. Yu, Attention and engagement aware multimodal conversational systems, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 593–597.
[24] M. Chen, S. Wang, P.P. Liang, T. Baltrušaitis, A. Zadeh, L.-P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM, 2017, pp. 163–171.

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335.
[26] R. Lowe, N. Pow, I. Serban, J. Pineau, The Ubuntu Dialogue Corpus: A Large dataset for Research in Unstructured Multi-Turn Dialogue Systems, arXiv:1506.08909 (2015).
[27] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised Learning of Universal Sentence Representations From Natural Language Inference Data, arXiv:1705.02364 (2017).
[28] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.
[29] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[30] A. Ritter, C. Cherry, B. Dolan, Unsupervised modeling of twitter conversations, in: Human Language Technologies: Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 172–180.
[31] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, in: Proceedings of the 2019 ACL, 2019, pp. 527–536.
[32] D. Lachowicz, pyenchant, (https://github.com/rfk/pyenchant) (Accessed in 2020).
[33] R. Ochshorn, M. Hawkins, Gentle: A Forced Aligner, 2016.
[34] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, ACM, 2010, pp. 1459–1462.
[35] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al., The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism, in: Proceedings of the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH 2013, 2013. Lyon, France.
[36] H. Xianyu, X. Li, W. Chen, F. Meng, J. Tian, M. Xu, L. Cai, SVR based double-scale regression for dynamic emotion prediction in music, in: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 549–553.
[37] M.-T. Luong, H. Pham, C.D. Manning, Effective Approaches to Attention-Based Neural Machine Translation, arXiv:1508.04025 (2015).
[38] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, X. Zhu, Commonsense knowledge aware conversation generation with graph attention., in: Proceedings of the 2018 IJCAI, 2018, pp. 4623–4629.
[39] M. Luong, E. Brevdo, R. Zhao, Neural Machine Translation (seq2seq) Tutorial, https://github.com/tensorflow/nmt (2017).
[40] C.-W. Liu, R. Lowe, I.V. Serban, M. Noseworthy, L. Charlin, J. Pineau, How Not to Evaluate your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation, arXiv:1603.08023 (2016).
[41] S. Liu, H. Chen, Z. Ren, Y. Feng, Q. Liu, D. Yin, Knowledge diffusion for neural dialogue generation, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 2018, pp. 1489–1498.
[42] M. Schröder, J. Trouvain, The German text-to-speech synthesis system MARY: a tool for research, development and teaching, in: Proceedings of the 4th ISCA Workshop on Speech Synthesis, 2001.
[43] J.A. Russell, A circumplex model of affect., J. Person. Soc. Psychol. 39 (6) (1980) 1161.

Tom Young got his Bachelor from Beijing Institute of Technology in 2018. Currently, he is a Ph.D.student under the supervision of Erik Cambria in the School of Computer Science and Engineering in Nanyang Technological University. His main research interests are dialogue systems, deep learning, and computer vision. Specifically, he applies memory-augmented neural networks to model human conversation. He is interested in expanding current chatbot systems to handle more complex environments with higher accuracy.



Vlad Pandelea received his Bachelor and Master of Science in Computer Science from Pisa University in 2017 and 2019, respectively. Since 2020, he is a PhD student at NTU under the supervision of Erik Cambria. His thesis focuses on the exploitation of multimodal information for dialogue systems. In addition to dialogue systems, his research interest lies in data analytics and in the application of deep learning techniques to a variety of fields, including sentiment analysis, time series and point processes.

Soujanya Poria received his B.Eng. in Computer Science from Jadavpur University (India) in 2013. In the same year, he received the best undergraduate thesis and researcher award and was awarded Gold Plated Silver medal from Jadavpur University and Tata Consultancy Service for his final year project during his undergraduate course. In 2017, Soujanya got his Ph.D. in Computing Science and Mathematics from the University of Stirling (UK) under the co-supervision of Amir Hussain and Erik Cambria. Soon after, he joined Nanyang Technological University as a Research Scientist in the School of Computer Science and Engineering. Later in 2019, he joined Singapore University of Technology and Design (SUTD), where he is now conducting research on aspect-based sentiment analysis in multiple domains and different modalities as an Assistant Professor.

Erik Cambria is the Founder of *SenticNet*, a Singapore-based company offering B2B sentiment analysis services, and an Associate Professor at *NTU*, where he also holds the appointment of Provost Chair in Computer Science and Engineering. Prior to joining NTU, he worked at Microsoft Research Asia and HP Labs India and earned his Ph.D. through a *joint programme* between the University of Stirling and MIT Media Lab. He is recipient of many awards, e.g., the 2018 *AI's 10 to Watch* and the 2019 *IEEE Outstanding Early Career* award, and is often featured in the news, e.g., Forbes. He is Associate Editor of several journals, e.g., NEUCOM, INFFUS, KBS, IEEE CIM and IEEE Intelligent Systems (where he manages the Department of *Affective Computing and Sentiment Analysis*), and is involved in many international conferences as PC member, program chair, and speaker.