Inferring User Emotive State Changes in Realistic Human-Computer Conversational Dialogs

Runnan Li
Tsinghua University
lirn15@mails.tsinghua.edu.cn

Zhiyong Wu
Tsinghua University
zywu@sz.tsinghua.edu.cn

Jia Jia*
Tsinghua University
jjia@tsinghua.edu.cn

Jingbei Li
Tsinghua University
jb-li15@mails.tsinghua.edu.cn

Wei Chen
Sogou, Inc.
chenweibj8871@sogou-inc.com

Helen Meng
The Chinese University of Hong Kong
hmmeng@se.cuhk.edu.hk

ABSTRACT

Human-computer conversational interactions are increasingly pervasive in real-world applications, such as chatbots and virtual assistants. The user experience can be enhanced through affective design of such conversational dialogs, especially in enabling the computer to understand the emotive state in the user's input, and to generate an appropriate system response within the dialog turn. Such a system response may further influence the user's emotive state in the subsequent dialog turn. In this paper, we focus on the change in the user's emotive states in adjacent dialog turns, to which we refer as user emotive state change. We propose a multi-modal, multi-task deep learning framework to infer the user's emotive states and emotive state changes simultaneously. Multi-task learning convolution fusion auto-encoder is applied to fuse the acoustic and textual features to generate a robust representation of the user's input. Long-short term memory recurrent auto-encoder is employed to extract features of system responses at the sentence-level to better capture factors affecting user emotive states. Multi-task learned structured output layer is adopted to model the dependency of user emotive state change, conditioned upon the user input's emotive states and system response in current dialog turn. Experimental results demonstrate the effectiveness of the proposed method.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Discourse, dialogue and pragmatics**; **Neural networks**; • **Human-centered computing** → *Human computer interaction (HCI)*;

KEYWORDS

human-computer interaction, emotive state changes prediction, multi-modal multi-task deep learning, convolution fusion auto-encoder, recurrent auto-encoder, structured output layer

*corresponding author.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240575

ACM Reference Format:
Runnan Li, Zhiyong Wu, Jia Jia, Jingbei Li, Wei Chen, and Helen Meng. 2018. Inferring User Emotive State Changes in Realistic Human-Computer Conversational Dialogs. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240575

1 INTRODUCTION

Automatic spoken dialog systems, known as interactive speech agents, receive speech as input and response via natural language to provide interactive services. With the development of intelligent speech interaction technologies, human-computer conversational interactions are increasingly pervasive in real-world applications, including virtual assistants and chatbots such as Apple Siri, Amazon Alexa, Microsoft Cortana and XiaoIce. A report[1] from Microsoft indicates Cortana has attracted 141 million monthly usages.

Human-human spoken interactions are highly expressive, using different features of emotions, intonations and styles to convey the underlying intent of the message [30]. To further enhance user experience, intelligent interaction systems are expected to capture and model such complex features for better understanding users' intentions, especially understanding the emotive state in the user's input [13, 33]. Automatic emotion recognition is thus becoming a focus in human-computer interaction research field. Emotion, as a key component in human cognition and communication processes, is embedded in the acoustic speech and the related transcribed text [23]. Text-based sentiment analysis has been developed and obtained remarkable achievements [26]. To better exploit the rich emotion-related information embodied in speech, acoustic-based emotion analysis is also proposed [9, 25]. [22] proposed a multi-modal fusion strategy in emotion recognition, gaining a significant performance improvement by utilizing the lexical, acoustic, visual features as well as their correlations simultaneously.

In communication process, participant's emotive state changes dynamically. To enhance human-computer communication experience, speech interaction systems are required to generate affective responses rather than neutral sentences [30]. Hence, a guide is needed for the systems to evaluate whether the response is appropriate or not. For example, to help an anxious user calm down and feel better, the systems should be able to gauge the influence of their responses to the user, respond as needed over the course of the spoken interactions and further provide appropriate and satisfactory interactions. In understanding what user is trying to tell,

[1] https://goo.gl/yGu2Ef

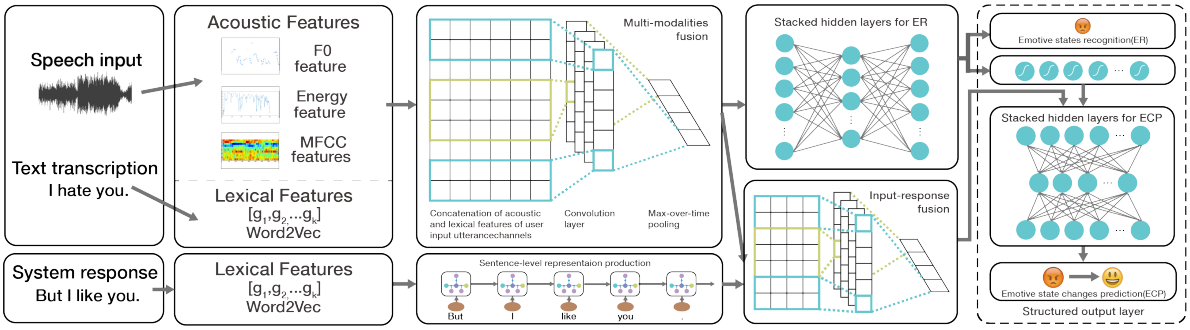

Figure 1: Overview of the proposed framework for inferring changes of user emotive states in conversational dialogs.

either verbally or nonverbally, and then inferring the user emotive state changes caused by responses, we can introduce the emotional intelligence into the speech interaction systems to further enhance user experiences.

The emotive state changes in dialog conversation can be defined as the difference of user emotive states between two adjacent dialog turns. To analyze the properties of emotive state changes, we constructed a real-world speech interaction dialog data set with the help of Sogou Speech Assistant. Some interesting observations are noticed after analyzing the collected data: emotive state changes exist in the real-world human-computer interaction dialog scenarios, triggered by system responses and related to the current states of user emotion. These findings inspired us to develop a framework to infer user emotive state changes using the user input utterances and system responses in dialog turns. For exploiting the dependence between emotive state changes and current user emotive states, a conditional structure is further employed.

The proposed framework for user emotive state changes inferring is illustrated in Fig.1. Related to different abstraction level, acoustic features and lexical features extracted from user input utterance are represented on a variety of formats and time resolution. Integrating the information with different modalities is not trivial and can significantly influence the effectiveness of the proposed framework. The influence caused by system responses to user is also necessary to be considered. Furthermore, as the user usually reacts to the system responses based on the understanding of the whole sentence, the framework should capture factors affecting user emotive states at the sentence-level of system responses. In this work, we introduce the use of $MTCF_{encoder}$ from multi-task learning convolution fusion auto-encoder (MTCF AE) to integrate acoustic features and lexical features from user input at feature-level. Developed from convolutional neutral network (CNN) fusion model [7] and multi-task auto-encoder (MT AE) [8], $MTCF_{encoder}$ in MTCF AE inherits the strong generalization performance of conventional multi-task learning while integrating signals at different layers through convolution and pooling. A long short term memory (LSTM) recurrent auto-encoder (RAE) [28] is trained to map system responses into sentence-level vector representation via $LSTM_{encoder}$, as illustrated by the sentence-level representation production component. Then in input-response fusion component, the convolution fusion (CF) approach [7] is further employed to integrate the encoded representations from $MTCF_{encoder}$ and $LSTM_{encoder}$ for predicting the targeted user emotive state changes. Considering the user emotive state changes is highly related to current user emotive states in dialog turns, a multi-task learning structured output layer (SOL) [29] is proposed to model the dependencies, where the prediction of user emotive state changes is conditioned upon the recognition result of user emotive states. Experimental results on public emotion database IEMOCAP and realistic interaction database have demonstrated the effectiveness of the proposed framework.

The main contributions of this paper lie in three aspects: 1) this is the first attempt to investigate the interactive influence of the system-generated response on the user's emotive state changes in human-computer interactive dialogs; 2) a novel fusion structure is suggested for multi-modal information integrating while reducing the time resolution; 3) an effective structure is suggested to infer user emotive states and emotive state changes simultaneously.

2 PROBLEM FORMULATION

Known. For conversations collected from realistic human-computer speech interaction dialog system, each conversation contains several dialog turns. In each dialog turn, user input one speech utterance and system will feedback one automatically generated response. For the given set of conversations C, each conversation $c_m \in C$ contains a set of dialog turns D_m. For one dialog turn $d_i \in D_m$, it consists of user input utterance $u_i = \{ua_i, ut_i\}$ and system response $r_i = \{rt_i\}$. For the user input utterance $u_i = \{ua_i, ut_i\}$, $ua_i = \left\{ua_i^1, ua_i^2, ..., ua_i^{F_i}\right\}$ and $ut_i = \left\{ut_i^1, ut_i^2, ..., ut_i^{W_i}\right\}$ are the set of acoustic features extracted at frame-level from speech input and lexical features at word-level from transcribed text. Specially, acoustic features ua_i and lexical features ut_i are aligned by exploiting a hidden Markov model (HMM) based forced alignment model, thus the ut_i is represented as $ut_i = \left\{ut_i^1, ut_i^2, ..., ut_i^{F_i}\right\}$ by upsampling lexical features to match the length of acoustic features, where F_i is the amount of frames of acoustic features ua_i. For each frame, ua_i^j $(1 \le j \le F_i)$ in ua_i indicates a D_{ua} dimensional vector representing various acoustic features (e.g. fundamental frequency, energy and spectral parameters), and ut_i^j in ut_i indicates a D_{ut}

dimensional word embeddings representing the lexical information. For the system response $r_i = \{rt_i\}$ with K_i words, lexical features $rt_i = \left\{rt_i^1, rt_i^2, ..., rt_i^{K_i}\right\}$ are extracted and represented using D_{rt} dimensional word embeddings. Specially, we concentrate on the influence caused by lexical content of system responses in this work, the influence caused by synthesized speech of the system response will be further discussed in the future work.

Definition 1. Emotive states. We employ the numerical dimensional Pleasure-Arousal-Dominance (PAD) [18] emotion model to describe emotive state in this work: 1) Pleasure-Displeasure Scale (P) measures human perceived level at pleasant or unpleasant; 2) Arousal-Nonarousal Scale (A) measures human perceived level at energized or soporific; 3) Dominance-Submissiveness Scale (D) represents the controlling and dominant versus controlled or submissive by human perceiving. Considering the dominance (D) dimension in PAD is highly related to the expression of emotion for dialog acts in spoken dialog interaction [32], the PAD model is adopted in this work for better describing the user emotive states. For given u_i in a dialog turn, its emotive state is denoted as $E_i = \left\{E_i^P, E_i^A, E_i^D\right\}$, where E_i^P, E_i^A, E_i^D are the quantized numerical levels at pleasure, arousal, and dominance dimensions respectively.

Definition 2. Emotive state changes. The changes of emotive state in this paper are defined as the difference of user's emotive PAD states in adjacent dialog turns. This definition is based on two assumptions: 1) user emotion in one conversation is continuous; 2) the user emotive state changes are triggered by dialog system responses while external environment is stable in a short period. For given contiguous user input utterance u_i and u_{i+1} labeled with E_i and E_{i+1}, we denote emotive state changes as $EC_i = \left\{EC_i^P, EC_i^A, EC_i^D\right\}$, where EC_i^P is calculated by $(E_{i+1}^P - E_i^P)$, EC_i^A is calculated by $(E_{i+1}^A - E_i^A)$, EC_i^D is calculated by $(E_{i+1}^D - E_i^D)$.

Problem. For an given user input utterance and its corresponding system response, automatically recognize the user emotive state of the current user input utterance and infer the emotive state change triggered by the system response. In this work, the proposed multi-modal multi-task learning framework is implemented to address the emotive state recognition task and emotive state changes prediction task simultaneously:

$$\{u_i, r_i\} = \{ua_i, ut_i, rt_i\} \rightarrow \{E_i, EC_i\}$$

3 DATA AND OBSERVATIONS

Data Collection. Being a data-driven task, the data used in this work is crucial. To ensure the reliability and reality of data, we collected raw data with Sogou Voice Assistant, a leading automatic speech dialog smart phone application in China. The collected raw data contains 4,052,847 Mandarin utterances from 221,459 users, the lexical information of each is provided by an automatic speech recognition (ASR) system with 5.5% word error rate (WER).

Preprocessing. Three different types of interactions are included in raw data: 'search', 'chat', and 'others' ('others' contains interactions like application launching etc.). The proportion of these three types of interactions are 48.62%, 34.65%, 16.73% respectively. The responses of 'search' and 'others' are normally fixed text or specific system operations, while the responses of 'chat' are automatically generated text or speech. Concentrate on conversational dialog system, interactions labeled with 'chat' are selected for further processing. The selected data contains 1,404,399 user input utterances, each utterance has one corresponding system response. We assume utterances from the same user with a short interval period, from seconds to half minute, can form an individual conversation. In addition, as research on dynamic changes requires more temporal information, conversations containing more than three utterances are selected. Hence, we collected a data set containing 98,376 conversations, each with 4 to 49 utterances.

Labeling. We randomly selected 2,000 user input utterances from the data set, and invited 3 human annotators to annotate these utterances with 'emotional' or 'neutral'. When annotators had opposite opinions, they stopped and had a discussion until they could achieve an agreement; if could not, this utterance would be abandoned. As a result, 1,125 utterances are annotated as 'emotional', and 875 utterances are annotated as 'neutral'. While we expected the conversations are more expressive than neutral, a simple deep neural network (DNN) based emotion detection model is implemented to further filter the data set. This model is trained to figure out whether the input utterance is 'neutral' or 'emotional' as a two-classification problem. Trained and tested on aforementioned labeled data, the filter network achieves 76.79% Recall and 87.16% Precision on the classification task. We then used this network to automatically annotate the utterances to find out the conversations with higher expression: all the utterances in such conversations are labeled as 'emotional'. In this way, 9,389 conversions with 52,064 utterances are selected.

For dimensional emotion annotating, the specification of labeling strategy of PAD [14] emotion model is followed. 15 annotators (five groups with each containing three annotators) and one expert were invited to describe the emotion of utterances using a 15-term simplified questionnaire [14]. Annotators were pre-trained for using the PAD questionnaire and the expert was asked to randomly check the annotations to ensure labeling quality. We finally annotated **1,124** conversation with **6,185** utterances, with **0.63** and **0.51** inter-annotator agreement on emotive states and emotive state changes respectively (elaborated in section 5.2 and Table.2).

Observations. We further analyzed the annotated data to investigate emotive state changes in interaction dialogs. The statistical results of annotation are shown in Table.1. As can be seen, for each discrete emotion category (i.e. angry(ANG), disgust(DIS), happy(HAP), sad(SAD), relax(REL),weak(WEA) and neutral (NEU)), the distribution of the annotated emotive state values confirms to the assumption of the PAD emotion model as well as the discrete emotion distribution in PAD emotional space as described in [14]. For example, the utterances with happy (HAP) emotion category possess all positive annotated P,A and D values, while the utterances with sad (SAD) emotion category are annotated with all negative P,A and D values. These results confirm the validity and effectiveness of our emotive state annotation procedures. The annotated data can thus be used to evaluate the proposed framework.

Observation 1: Emotive state changes exist and are triggered by responses. In dialog turns, user emotive states are changing dynamically. As the example shown in Fig.2, user was in negative mood at the beginning, but became positive with the influence of appropriate affective system responses. In this case, system responses show

Table 1: Average PAD annotation values for different emotions and the corresponding emotive state changes.

	ANG	DIS	FEA	HAP	NEU	SAD	REL	WEA
P	**-0.31**	**-0.28**	**-0.17**	**0.40**	**0.34**	**-0.26**	**0.33**	**0.41**
A	**0.47**	**-0.18**	**0.37**	**0.47**	**-0.15**	**-0.32**	**-0.12**	**0.42**
D	**0.48**	**0.41**	**-0.09**	**0.39**	**-0.17**	**-0.23**	**0.33**	**-0.11**
EC^P	**0.11**	**0.06**	**0.08**	**-0.17**	**-0.11**	**0.10**	**-0.12**	**-0.21**
EC^A	**-0.10**	**0.20**	**-0.09**	**-0.12**	**0.18**	**0.24**	**0.25**	**-0.15**
EC^D	**-0.14**	**-0.09**	**0.15**	**-0.10**	**0.13**	**0.26**	**-0.07**	**0.17**

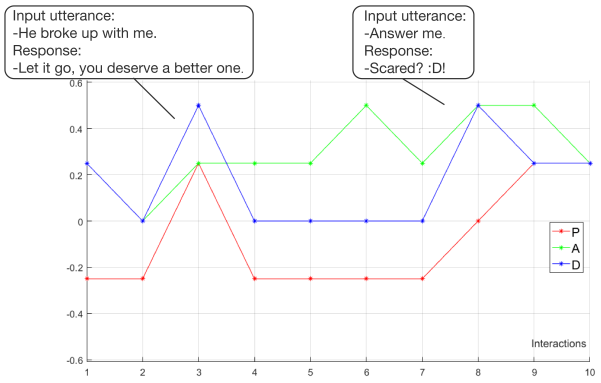

Figure 2: An example of user emotive state changes in dialog turns, the system responses had comforted user's mood.

their effectiveness in comforting user. However, in some other case, user was even infuriated by unsuitable system responses.

Observation 2: Emotive state changes are related to the current emotive states. As Table.1 shows, for one specific emotion, the emotive state changes are statistically opposite to current emotive states, normally trending from emotional states to neutral. This observation indicates users can relieve from negative emotion in the interaction with automatic spoken dialog system, and also indicates current dialog system is still lack in increasing and sustaining the positive emotion of user.

Summarization. We find the user emotive state changes exist in real-world human-computer interaction dialog scenarios, triggered by system responses and related to the current states of user emotion. These findings inspire us to introduce a multi-modal multi-task framework to infer user emotive state changes conditioned upon user input and system response in human-computer communication dialog turns to further enhance user experiences.

4 METHODOLOGY

To infer user emotive state changes, following challenges are involved: 1) how to produce robust representation of user input utterances considering different modalities; 2) how to produce sentence-level representation of system responses and exploit it in user emotive state changes prediction; 3) how to model the dependency between user current emotive states and emotive state changes.

To solve aforementioned challenges, we propose the usage of a multi-modal multi-task learning framework to predict user emotive state changes in dialog turns. As illustrated in Fig.3, the proposed framework contains four components: 1) multi-modalities fusion component to produce robust representation of user input utterances using multi-task convolution fusion auto-encoder; 2) sentence-level representation production component to compress the system response using LSTM based recurrent auto-encoder; 3) input-response fusion component to produce conditioned representation upon user input and system response; 4) multi-task learning structured output layer (SOL) to infer user current emotive states and emotive state changes simultaneously while modeling the dependency between emotive states and emotive state changes.

Multi-modalities fusion. To fuse different but related modalities extracted from user input utterance into one robust representation, we introduce the usage of multi-task learning convolution fusion auto-encoder (MTCF AE). Inspired by [16], and developed from CNN fusion model [7] and multi-task auto-encoder (MT AE) [8], $MTCF_{encoder}$ in MTCF AE inherits the strong generalization performance of conventional multi-task learning while integrating signals at different layers through convolution and pooling.

This convolutional neutral networks based approach is constructed by stacking alternatively convolutional layers and pooling

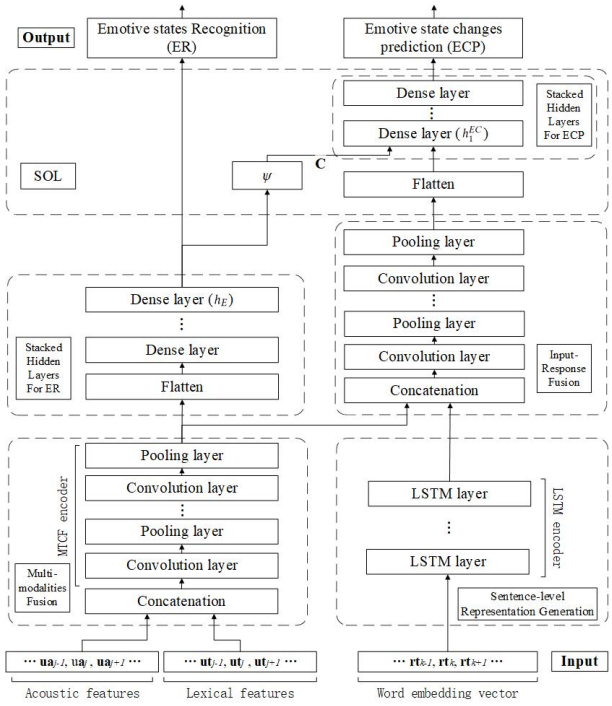

Figure 3: Overall structure of the proposed framework. The multi-modalities fusion component employs $MTCF_{encoder}$ to produce a robust representation of user input utterance, and the sentence-level representation production component employs $LSTM_{encoder}$ to compress the system response, the input-response component employed a convolution fusion approach to fuse the representations produced from user input utterance and system response, structured output layer (SOL) is employed to infer user emotive states and emotive state changes simultaneously.

layers. Each convolutional layer contains a pack of neurons processing sequentially consecutive patches on input, i.e. on the temporal dimension. Extracting at every position of input, each neuron produces one value to create a new signal referred to as feature map. Pooling layers are implemented by applying statistical function, i.e. average or maximum, on non-overlapping patches to reduce the dimensionality of feature maps. By stacking several convolutional and pooling layers, $MTCF_{encoder}$ can integrate different modalities and reduce the time resolution upon the temporal features extracted from user input utterances [15].

For acoustic features ua_i and aligned lexical features ut_i extracted from utterance u_i, the first step in $MTCF_{encoder}$ is to process frame-level concatenation:

$$u_i^{cat} = Concat(ua_i, ut_i) \tag{1}$$

Convolution layer is then convolved on the concatenated sequences with filters f and biases b:

$$h^k = \varphi(\sum_{l=L} h^l \oplus f^k + b^k) \tag{2}$$

where h^k is the latent representation of the k-th feature map of the current layer, φ is Rectified Linear Unit (ReLU) activation function, h^l is l-th feature map of feature maps group L of upper layer or l-th channel of input u_i^{cat} with total L channels for the first convolutional layer, $\oplus$ denotes the convolution operation. Max-pooling is used as pooling layer to reduce the time resolution. The output of top pooling layer is used as the fusion representation R_i of input utterance u_i.

We employ the unsupervised multi-task learning auto-encoder structure to train the $MTCF_{encoder}$ [8] as shown in Fig.4. In training, the original input features from different modalities are encoded by $MTCF_{encoder}$ and then be reconstructed by $MTCF_{decoder}$. Constructed with stacked upsampling layers and deconvolution layers, the $MTCF_{decoder}$ firstly restore the time resolution via upsampling and then reconstruct the input signals via deconvolution operations. By further employing multi-task training style, two types of reconstruction tasks are performed in training: 1) self-domain reconstruction of features and 2) between-domain reconstruction among modalities. The self-domain reconstruction ensures the information is maintained in encoding process, and the between-domain reconstruction guarantees the correlation information between modalities being further considered in encoding. These two types can thus help $MTCF_{encoder}$ generating a robust fusion representation of acoustic features and aligned lexical features.

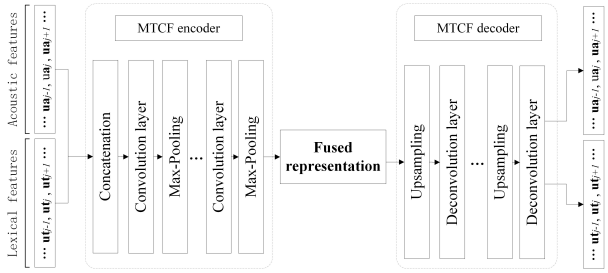

Figure 4: The framework of MTCF AE.

Sentence-level representation production. As the user in human-computer interaction usually reacts to the system responses based on the understanding of the entire sentence, the framework should capture factors affecting user emotive states at the sentence-level of the system responses. A LSTM based recurrent auto-encoder [28] is employed to produce the sentence-level representation of system responses. Similar to sequence-to-sequence generation tasks, this structure is implemented as shown in Fig.5: 1) converting the system response into a sentence-level vector representation via $LSTM_{encoder}$, 2) decoding the compressed vector representation to an output string of words via $LSTM_{decoder}$. Two separate LSTMs are implemented as encoder and decoder without considering sentence structures.

For lexical features rt_i of given system response r_i with K_i words, $LSTM_{encoder}$ computes the hidden vector sequence $h = (h_1, h_2, ..., h_{K_i})$ from $s = 1$ to K_i with following equations:

$$f_s = \sigma(W_f rt_{i,s} + U_f h_{s-1} + b_f) \tag{3}$$
$$i_s = \sigma(W_i rt_{i,s} + U_i h_{s-1} + b_i) \tag{4}$$
$$o_s = \sigma(W_o rt_{i,s} + U_o h_{s-1} + b_o) \tag{5}$$
$$c_s = f_s \circ c_{s-1} + i_s \circ tanh(W_c rt_{i,s} + U_c h_{s-1} + b_c) \tag{6}$$
$$h_s = o_s \circ tanh(c_s) \tag{7}$$

where σ is the Sigmoid activation function, f, i, o and c are the *input gate*, *forget gate*, *output gate* and *memory cell* activation vectors respectively, W, U and b items are the weight matrices and bias vectors of each gate. The hidden vector h_{K_i} is used as the sentence-level vector representation S_i of system response r_i. In decoding, S_i is used to replacing $rt_{i,s}$, from $s = 1$ to K_i, in Eq.(3)-(6), to reconstruct the original rt_i.

Input-response fusion. As user emotive state changes are based on user current input and conditioned upon system response, we propose the use of a convolution fusion (CF) component to integrate information from fused representation of user input utterances and sentence-level representation of system responses.

To fuse representation R_i of utterance u_i and sentence-level vector representation S_i of corresponding system response r_i, the concatenation is firstly processed at frame-level from $t = 1$ to T in R_i, where T is the time step length of R_i:

$$C_i^t = Concat(R_i^t, S_i) \tag{8}$$

the fused representation FS_i is then calculated following Eq.(2).

Emotive states recognition. With aforementioned components, a baseline emotive states recognition model can be implemented: it employs pre-trained $MTCF_{encoder}$ to produce fusion representation R_i of input utterance u_i, and then uses stacked non-linear

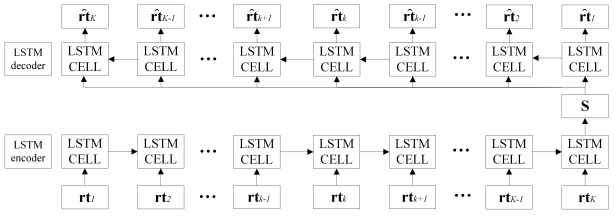

Figure 5: The framework of LSTM recurrent auto-encoder.

full-connection hidden layers $\{h_1^E, h_2^E, ..., h_L^E\}$ for wrapping to predict the PAD value, the first non-linear hidden layer and the emotion PAD prediction output is thus computed as:

$$h_1^E = \sigma(W_{R_i h_1^E} R_i + b_1^E) \tag{9}$$

$$O_E = \sigma_E(h_L^E) \tag{10}$$

where $\sigma, W, b, \sigma_E, O_E$ are the non-linear activation function, weights matrix, bias, linear output function and predicted emotion PAD state values E_i respectively.

Emotive state changes prediction. As user emotive state changes are highly related to user current emotive states, we propose the use of multi-task learning structured output layer (SOL) to simultaneously infer user emotive states E_i and emotive state changes EC_i. In conventional multi-task learning framework, emotive states recognition task and emotive state changes prediction task share the hiddens of aforementioned components. For emotive state changes prediction task, additional non-linear full-connection hidden layers $\{h_1^{EC}, h_2^{EC}, ..., h_L^{EC}\}$ using input-response fusion result FS_i as input are employed. The current emotive states is calculated following Eq.(9) and emotion changes is computed as:

$$h_1^{EC} = \sigma(W_{FS_i h_1^{EC}} FS_i + b_1^{EC}) \tag{11}$$

$$O_{EC} = \sigma_{EC}(h_L^{EC}) \tag{12}$$

where O_{EC} is the predicted emotive state changes EC_i.

Based on observed relationship between user emotive states and corresponding emotive state changes, SOL is suggested in this work for explicitly modeling the dependency of the primary emotive state changes prediction task on the auxiliary emotive states recognition task. This is realized by feeding the emotive states recognition task's hidden layer output h_E through an activation function $\Psi(\cdot)$, such as Tanh or ReLU, to model the correlation between the two tasks before being augmented to the hidden layer h_1^{EC} while the weight matrix $\mathbf{C}$ used to connect the two tasks is applied. The first layer h_1^{EC} of the hidden layers is thus modified as,

$$h_1^{EC} = \sigma(W_{FS_i h_1^{EC}} FS_i + b_1^{EC} + \mathbf{C}\Psi(h_L^E)) \tag{13}$$

Precursors of the same SOL structure have been previously studied for recurrent neural network language modeling for predicting morphologically decomposed stem and suffix features [2, 29].

In common with the conventional multi-task learning framework, models with SOL can be trained by minimizing a global cost function expressed as a weighted sum of the two task specific separate error costs. This is given by

$$F_g = \alpha F_{EC} + (1 - \alpha) F_E \tag{14}$$

where F_{EC} and F_E are the costs generated by the main task (emotive state changes prediction) and the auxiliary task (emotive states recognition) computed as mean squared errors (MSE), and α is a tunable weighting parameter adjusting the contributions from the main and auxiliary tasks.

Using SOL as the output layer inherits the strong generalization performance and robustness of conventional multi-task learning facilitated by shared hidden layers and joint training over multiple tasks [5]. The use of structured output layer further allows both the regularization properties of the comparatively simpler auxiliary task of emotive states recognition and its direct effect on the primary emotive state changes prediction task to be fully exploited.

5 EXPERIMENTS AND DISCUSSION

The proposed framework is evaluated in two experiments. In the first evaluation, we compare the proposed framework with other state-of-the-art emotion recognition approaches on the tasks of emotive states recognition and emotive state changes prediction using the public emotional database IEMOCAP [4] and the collected realistic conversational dialogs data set (RCD). In the second experiment, we investigate the performance as well as the contribution of individual components in the proposed framework.

5.1 Experimental Setup

Data set. Following aforementioned construction process, we established a realistic speech interaction benchmark data set RCD containing unlabeled 9,389 conversions with 52,064 utterances and corresponding system responses. We manually annotated 1,124 conversations with 6,185 utterances from the data set following PAD emotion model. The distribution as well as the discrete degree of annotated emotive states are satisfied with the assumption in [14] as elaborated in section 3.

The IEMOCAP database contains 12 hours of audio-visual conversations in English, categorized according to the emotion: anger, happiness, sadness,neutral, excitement, frustration, fear, surprise, and others. We form a four-class emotion classification dataset containing $\{happy, angry, sad, neutral\}$ for experiments following the experimental setup of state-of-the-art approaches, thus 5,531 utterances are involved.

Features. Acoustic features are extracted from raw user input utterances using LibROSA [17] speech toolkit with 25 msec frame window size and 10 msec frame intervals. Totally, 41-dimensional frame-level acoustic features are extracted, containing 39-dimensional Mel-frequency cepstral coefficients (MFCCs), 1-dimensional logarithmic fundamental frequency (log F0), and 1-dimensional energy. Lexical features are extracted from both user input utterances and system responses through two steps: 1) word segmentation on utterances with a Chinese lexical analyzer Thulac Tool [27]; 2) getting 300-dimensional vector representation of words using Word2Vec [19] model. All utterances and responses from raw data, about 4 million utterances and 1.4 million responses, are used to train the Word2Vec model. For IEMOCAP based experiments, the Word2Vec toolkit proposed in [19] is used to represent the lexical information of utterances. For features extracted from user input utterances, frame-level alignment between acoustic features and lexical features is processed with an HMM-based approach. Input features were normalized to the range of [-0.99, 0.99] and targeted labels, including emotive states and emotive state changes, were normalized to zero mean and unit variance.

Construction setting. $MTCF_{encoder}$ contains 4 convolutional layers, each of 32 1-D convolution filters with a window length of 3 and stride of 1. Each convolutional layer is followed with a 1-D max-pooling layer, which is applied with a stride of 2. $LSTM_{encoder}$ contains a stack of 2 uni-directional LSTM layers with 64 units. MTCF AE is pre-trained using all the unlabeled user input utterances and LSTM RAE is pre-trained using all the system responses in raw

Table 2: The performance of state-of-the-art approaches and the proposed approach on IEMOCAP and RCD databases. ER:emotive states recognition. ECP: emotive state changes prediction. UA, the higher the better. MAE, the lower the better. CCC, the higher the better. (*: This approach used acoustic features only. **:System 2(S2) in section 5.3 for emotive states recognition with $MTCF_{encoder}$ employed. *: System 6 (S6) in section 5.3 with all components employed)**

	Method	IEMOCAP ER (UA,%)	RCD ER (MAE)	RCD ER (CCC)	RCD ECP (MAE)	RCD ECP (CCC)
[APSIPA ASC, 2012]	SVM	67.4	0.37	0.31	0.42	0.01
[ICASSP, 2015]	SVM	69.2	0.33	0.35	0.41	0.01
[ICDM, 2016]	CNN	65.1	0.35	0.34	0.36	0.05
[ICASSP, 2017]*	RNN	58.8	0.37	0.29	0.44	0.01
[ACL, 2017]	LSTM	**75.6**	**0.30**	**0.44**	0.37	0.06
The proposed**	$MTCF_{encoder}$	74.8	0.31	0.42	-	-
The proposed***	Hybrid	-	0.32	0.41	**0.28**	**0.31**
Inter-annotator agreement		-	-	0.63	-	0.51

data. The input-response fusion component has the same structure as $MTCF_{encoder}$. We use $Tanh$ as activation function ψ in Eq.(13) and $\alpha = 0.6$ in Eq.(14) in the proposed system.

Implementation and training. As a general setting, we utilize stochastic optimization with a mini-batch size of 128 samples, Adam [10] algorithm is employed as optimizer with an initial learning rate at 10^{-4} with $4 * 10^{-6}$ decay over each update. The proposed framework and all the comparisons are implemented using Keras [6] deep learning framework with Tensorflow [1] as backend. Mean squared errors (MSE) is used as the loss function of emotive states recognition task and emotive state changes prediction task. For models with LSTM components, back-propagation through time [31] is employed for training. The performance of implemented approaches are assessed by mean absolute error (MAE) and lin's concordance correlation coefficient (CCC) [12]. All the experimental results reported in this paper were based on 5-fold cross validation[11].

5.2 Comparison to state-of-the-art

In this experiment, we firstly evaluate the emotive states recognition model in the proposed framework on IEMOCAP database comparing to state-of-the-art approaches. We also employ these approaches to compare with the proposed framework using RCD dataset on both emotive states recognition task and emotive state changes prediction task.

Specially, when evaluated on IEMOCAP database, the loss function of the proposed emotive states recognition model is changed to cross-entropy. When evaluated on RCD, support vector regression (SVR) [3] is employed in SVM based approaches. For inferring user emotive state changes, we simply change the prediction target of state-of-the-art approaches from emotive states labels to emotive state changes labels without modifying any structure.

Comparison methods. We collected state-of-the-art approaches with reported experimental result on IEMOCAP to compare with the proposed framework:

(1) [APSIPA ASC, 2012] using support vector machine (SVM) and decision trees based classifiers to address the sentence-level multi-modal emotion recognition problem [24].
(2) [ICASSP, 2015] using both early fusion and late fusion for sentence-level acoustic and lexical features to recognize emotion with SVM based classifier [9].
(3) [ICDM, 2016] using CNN to extract features from multi-modalities and a multiple kernel learning classifier to recognize emotion[22].
(4) [ICASSP, 2017] using recurrent neural network (RNN) and local attention based feature pooling strategy to produce emotionally relevant features for emotion recognition [20].
(5) [ACL, 2017] using a LSTM based model to capture contextual information between utterance-level features to recognize emotion [21].

The evaluation metrics employed for IEMOCAP based experiments is unweighted accuracy (UA), the mean of accuracies for different emotion categories [24], that is commonly used in aforementioned state-of-the-art approaches.

Experiment result. As shown in Table.2, the proposed framework has achieved an on par performance with state-of-the-art approaches on emotive states recognition (ER) task when evaluated with IEMOCAP and RCD databases. For emotive state changes prediction (ECP) task, state-of-the-art approaches achieved limited performance, while the proposed approach gained significant improvement. Even for humans, the perception of emotive states and changes are unstable, which can be validated by CCC, the inter-annotator agreement on emotive states and emotive state changes are 0.63 and 0.51, respectively.

5.3 Contribution of individual components

Comparison systems. In this experiment, we evaluate the performance as well as the contribution of each individual component employed in the framework. Five comparison systems with different combination of components are implemented to compare with the proposed framework. Specially, components employed in the comparisons have the same structure with the proposed framework. However, the numbers of filters/units of convolution/dense layers are proportionally balanced to ensure the total parameters employed by the models are on par with the proposed framework.

(1) system 1 (S1), baseline emotive states recognition model, features extracted from difference modalities are simply concatenated as input for the system.
(2) system 2 (S2), $MTCF_{encoder}$ is employed to fuse modalities from user input utterances for emotive states recognition;

Table 3: Experimental results with different combination of components. CF: convolution fusion.

	Parameters	Modalities Fusion	System response Representation	Input-response Fusion	Multi-task Learning	E^{PAD} MAE	E^{PAD} CCC	EC^{PAD} MAE	EC^{PAD} CCC
S1	6.46M	*Concatenation*	NO	NO	NO	0.38	0.11	N/A	N/A
S2	6.47M	$MTCF_{encoder}$	NO	NO	NO	**0.31**	**0.42**	N/A	N/A
S3	6.53M	$MTCF_{encoder}$	NO	NO	NO	N/A	N/A	0.37	0.04
S4	6.71M	$MTCF_{encoder}$	$LSTM_{encoder}$	CF	NO	N/A	N/A	0.32	0.11
S5	6.21M	$MTCF_{encoder}$	$LSTM_{encoder}$	CF	YES	0.32	0.40	0.29	0.22
S6	6.31M	$MTCF_{encoder}$	$LSTM_{encoder}$	CF	SOL	0.32	0.41	**0.28**	**0.31**

(3) system 3 (S3), baseline single-task learning emotive state changes prediction model using $MTCF_{encoder}$ for integrating modalities from user input utterances only (sharing the same structure with S2);
(4) system 4 (S4), single-task learning emotive state changes prediction model employing $LSTM_{encoder}$ to produce sentence-level representation of system response and input-response fusion component (CF) to integrate information from both user input utterances and system response;
(5) system 5 (S5), conventional multi-task learning method inferring user emotive states and emotive state changes simultaneously;
(6) system 6 (S6), the proposed multi-task learning model with structured output layer (SOL) where the primary user emotive state changes prediction task is conditioned on the auxiliary emotive states recognition task.

Experimental result. As illustrated in Table.3, system 1 (S1) has achieved a limited performance in emotive states recognition task. By using $MTCF_{encoder}$ as the multi-modalities fusion component in S2, the performance has gained significant improvement from 0.38 to 0.31 (-18.4%) in MAE and from 0.11 to 0.42 (+281%) in CCC on emotive states recognition task. When targeting on emotive state changes prediction task, S3 has achieved limited performance. By using $LSTM_{encoder}$ as sentence-level representation production component for system response and input-response component (CF) to consider information from both user input and system response, S4 has gained an improvement from 0.37 to 0.32 (-13.5%) in MAE and from 0.04 to 0.11 (+175%) in CCC on emotive state changes prediction task. By modifying to multi-task learning style, S5 achieves appropriate performance in emotive state changes prediction task, from 0.32 to 0.29 (-9.3%) in MAE and from 0.11 to 0.22 (+100%) in CCC. By employing SOL, S6 has gained a further improvement comparing to S5, from 0.29 to 0.28 (-3.5%) in MAE and from 0.22 to 0.31 (+40%) in CCC.

5.4 Discussion

As shown in the experimental results, evaluated with the realistic interaction data, the proposed framework has outperformed state-of-the-art approaches on user emotive state changes inferring task. Comparing to state-of-the-art approaches, such performance gains of the proposed framework come from the following aspects:

(1) $MTCF_{encoder}$ is employed to fuse the acoustic and lexical features from user input utterances, which can improve the performance of emotive states recognition and provide robust representation of user input utterances for following emotive state changes prediction.
(2) Sentence-level representation of system responses is generated by $LSTM_{encoder}$, and further fused with the representation of user input using convolution fusion component. The combination of these components can help the proposed framework to infer user emotive changes considering both user input utterances and the influence of system responses. Experimental results have indicated the importance and effectiveness of these components.
(3) With the use of structure output layer (SOL), the proposed framework can utilize the dependency of user current emotive states and possible emotive state changes in inference, which can further improve the overall performance of the proposed framework.

6 CONCLUSION

To figure out the emotive state changes caused by responses in automatic speech dialog systems, we presented a multi-modal multi-task learning framework to infer user emotive states and emotive state changes simultaneously. Multi-task learning convolution fusion auto-encoder is applied to fuse the acoustic and textual features to produce a robust representation of user's input. Long-short term memory recurrent auto-encoder is employed to produce sentence-level representation of system responses to better capture factors affecting user emotive states. Multi-task learned structured output layer is adopted to model the dependency of user emotive state changes upon the user emotive states in current dialog turn. Experimental results on public emotional database IEMOCAP and real-world interaction database have illustrated the effectiveness of the proposed framework in predicting user emotive state changes.

7 ACKNOWLEDGMENT

This work is supported by National Key Research and Development Plan (2016YFB1001200), the Innovation Method Fund of China (2016IM010200), National Natural Science Foundation of China (NSFC) - Research Grant Council of Hong Kong (RGC) joint fund (61531166002, N_CUHK404/15), NSFC (61433018, 61375027), National Social Science Foundation of China (13&ZD189). We would also like to thank Sogou Inc. and Tiangong Institute for Intelligent Computing, Tsinghua University for their support.

REFERENCES

[1] Martín Abadi and Ashish Agarwal et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015). https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Ebru Arisoy and Murat Saraclar. 2016. Compositional Neural Network Language Models for Agglutinative Languages.. In *INTERSPEECH*. 3494–3498.

[3] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. 2007. Support vector regression. *Neural Information Processing-Letters and Reviews* 11, 10 (2007), 203–224.

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

[5] Rich Caruana. 1998. Multitask learning. In *Learning to learn*. Springer, 95–133.

[6] François Chollet et al. 2015. Keras. https://github.com/fchollet/keras. (2015).

[7] Christoph Feichtenhofer, Axel Pinz, and AP Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. (2016).

[8] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*. 2551–2559.

[9] Qin Jin, Chengxin Li, Shizhe Chen, and Huimin Wu. 2015. Speech emotion recognition with acoustic and lexical features. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 4749–4753.

[10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[11] Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. Montreal, Canada, 1137–1145.

[12] I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* (1989), 255–268.

[13] Chul Min Lee and Shrikanth S Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing* 13, 2 (2005), 293–303.

[14] Xiaoming Li, Haotian Zhou, Shengzun Song, Tian Ran, and Xiaolan Fu. 2005. The reliability and validity of the Chinese version of abbreviated PAD emotion scales. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 513–518.

[15] Héctor P Martínez and Georgios N Yannakakis. 2014. Deep multimodal fusion: Combining discrete events and continuous signals. In *Proceedings of the 16th International conference on multimodal interaction*. ACM, 34–41.

[16] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*. Springer, 52–59.

[17] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*. 18–25.

[18] Albert Mehrabian. 1980. *Basic Dimensions for a General Psychological Theory Implications for Personality, Social, Environmental, and Developmental Studies*.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[20] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2227–2231.

[21] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 873–883.

[22] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 439–448.

[23] K Sreenivasa Rao, Tummala Pavan Kumar, Kusam Anusha, Bathina Leela, Ingilela Bhavana, and SVSK Gowtham. 2012. Emotion recognition from speech. *International Journal of Computer Science and Information Technologies* 3, 2 (2012), 3603–3607.

[24] Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad. 2012. Ensemble of svm trees for multimodal emotion recognition. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 1–4.

[25] Björn Schuller, Gerhard Rigoll, and Manfred Lang. 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, Vol. 1. IEEE, I–577.

[26] Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*. ACM, 1556–1560.

[27] Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. *Thulac: An efficient lexical analyzer for chinese*. Technical Report. Technical Report.

[28] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[29] Pawel Swietojanski, Peter Bell, and Steve Renals. 2015. Structured output layer with auxiliary targets for context-dependent acoustic modelling. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[30] Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Bonin, et al. 2015. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation* 7, 4 (2015), 397–413.

[31] Paul J Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78, 10 (1990), 1550–1560.

[32] Zhiyong Wu, Helen M Meng, Hongwu Yang, and Lianhong Cai. 2009. Modeling the expressivity of input text semantics for Chinese text-to-speech synthesis in a spoken dialog system. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 8 (2009), 1567–1576.

[33] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. IEEE* 101, 5 (2013), 1160–1179.