PREDICTING DIALOGUE SUCCESS, NATURALNESS, AND LENGTH WITH ACOUSTIC FEATURES

Alexandros Papangelis[1], Margarita Kotti[1], Yannis Stylianou[1,2]

[1]Speech Technology Group, Toshiba Research Europe Ltd., UK
[2]Department of Computer Science, University of Crete, Greece

ABSTRACT

Statistical methods for Spoken Dialogue Systems have been shown to reduce the cost of development, while successfully handling a variety of applications. However, such systems are usually trained with simulated users or paid subjects in controlled settings. While this may be sufficient to jump-start learning in the various sub-components, learning is very much dependent on the complete knowledge that we have about the interaction. Relatively few works have focused on this problem, and we here propose to extract low-level audio descriptors and use them as input to various classifiers, namely support vector machines, Gaussian process regressors, and random forests, to predict metrics that are constituents of user satisfaction from acoustic features. While our approach is not directly comparable to the current state of the art, results show that models using the proposed feature set outperform models that use state of the art features extracted from the belief state.

Index Terms— spoken dialogue management, acoustic features, dialogue quality, user satisfaction

1. INTRODUCTION

Following the recent advances in Statistical Spoken Dialogue Systems (SSDS) [1, 2, 3, 4, 5, 6, 7, e.g.], the need for hand-tuning the various components has largely been alleviated, provided that labeled data are available for each component. End-to-end SSDS [8, 9, 10], on the other hand, do not need labels for each sub-component, but still require some overall measure of dialogue quality. Accordingly, the aim of this work is to propose an innovative way to measure the dialogue quality directly from the user's speech signal.

Traditionally, data collection for SSDS is performed in controlled environments, where the ground truth is always available and can be used to train and assess the system's performance under this "constructed" partial observability. Specifically, information about the true intentions of the user (i.e. the user's goal) is usually predetermined, and dialogue success is judged based on this premise. However, outside this controlled environment, it is impossible to unobtrusively retrieve information about the user's goal, which may possibly be changing over time. Therefore, assessing dialogue success is not trivial in real-world applications. Prior work has proposed various approaches to this problem (presented in the next section), however, to the best of our knowledge, no work has explored the use of acoustic features to predict metrics of dialogue quality. We propose to use simple acoustic features to predict dialogue success, naturalness, and length of the interaction, and we show that we achieve higher accuracy rates than when using the best performing feature set that was proposed in the literature for data similar to ours.

1.1. Prior Work

A limited number of studies appear in the literature that take acoustic features into account to improve the quality of SSDS. For example, the system that appears in [11] takes advantage of pitch, energy and duration features that are given as input to Support Vector Machines (SVMs) so as to discriminate between 8 dialogue acts. The ultimate aim is to understand the structure of spoken language. The same classifier is also used in [12] in order to detect reported speech in dialogue systems. Specifically, prosodic and timing features are used to analyse phone conversations aiming to understand their structure. In an additional work [13], classification trees were used in order to detect sarcasm based on prosodic and spectral features, aiming to create a dialogue agent that is able to understand sarcasm. Finally, phonetic distances have been used in [14]. In that case, SVM as well as Repeated Incremental Pruning to Produce Error Reduction Repetitions (RIPPER) classifiers are used. The motivation of the authors is that repetition can be a symptom of problematic communication between users and systems. In all those works, the speech signal has been used to augment the SSDS's efficiency. [15] propose a classifier for problematic or normal dialogues using multiple types of features but not directly derived from the speech signal. [16] use acoustic features to predict the type of anger (hot or calm), but not metrics of dialogue quality, as proposed in this paper. To the best of the authors' knowledge, this is the first attempt to use features from the speech signal to predict metrics relevant to dialogue quality, that can subsequently be used to improve the dialogue experience.

Prior work has also suggested frameworks and metrics that correlate with user satisfaction, which most often is what SSDS really want to optimise. PARADISE [17] is a very well known example of frameworks that suggest a number of metrics related to user satisfaction, such as: task success, dialogue length, ratio of system's inappropriate actions, etc. [17] also propose to estimate dialogue success based on the k coefficient using labelled data. [18] propose a reward inference scheme, called Task Completion Transfer Learning, to estimate dialogue quality on-line and evaluate it on simulated dialogues. In order to classify a dialogue as successful or not in dialogues with humans, [19] and [20] propose to use Recurrent Neural Networks trained with an agenda-based simulated user. This approach was shown to have potential to generalise to interactions with human users, and since it is the closest work to our approach, we here regard it as the state of the art.

In this paper, we propose to use simple acoustic features that can be easily extracted from real-world interactions. We aim to predict metrics related to dialogue success as well as naturalness of interaction that indicate user satisfaction [17, 21], and can be used to train dialogue policies or end-to-end SSDS. The rest of the paper is organised as follows. In the next section we briefly introduce SSDS and dialogue management in particular, in section 3 we present our

System: *Hello! I'm here to help you find a suitable laptop. Now please tell me your requirements.*
User: *I want a laptop for business.*
S: *What's the price range you have in mind for the laptop? Such as a cheap laptop, an expensive laptop, or a moderately priced laptop?*
U: *I don't care.*
S: *Satellite u920t-117 is a very good computer for business computing.*
U: *Is that in the mid-weight range?*
S: *The weight of Satellite u920t-117 is 1.52 kilograms.*
...

Table 1. Excerpt of the transcript of an interaction between an AMT worker and our SSDS.

dataset and discuss data pre-processing, in the following section we present our experiments and results and in section 5 we conclude.

2. STATISTICAL SPOKEN DIALOGUE SYSTEMS

Partially Observable Markov Decision Processes (POMDP) [22] have been preferred in dialogue management due to their ability to handle uncertainty, which is inherent in human communication. A POMDP Dialogue Manager (DM) typically receives an n-best list of language understanding hypotheses, which are used to update the belief state (reflecting an estimate of the user's goals). Using Reinforcement Learning (RL), the system selects a response that maximises the long-term return of the system. This response is typically selected from an abstract action space and has to be converted to text through language generation. Concretely, a POMDP is defined as a tuple $\{S, A, T, O, \Omega, R, \gamma\}$, where S is the state space, A is the action space, $T : S \times A \rightarrow S$ is the transition function, $O : S \times A \rightarrow \Omega$ is the observation function, Ω is a set of observations, $R : S \times A \rightarrow \Re$ is the reward function and $\gamma \in [0, 1]$ is a discount factor of the expected cumulative rewards $J = E[\sum_t \gamma^t R(s_t, a_t)]$. A policy $\pi : S \rightarrow A$ dictates which action to take from each state. An optimal policy $\pi^\star$ selects an action that maximises the expected returns of the POMDP, J. Learning in RL consists exactly of finding such optimal policies; however, due to state-action space dimensionality, approximation methods [23, e.g.] need to be used for practical applications.

Moreover, the definition of the reward function is crucial to learning, as it dictates the optimality of policies. Typical reward functions for SSDS are of the form:

$$R(s, a) = \begin{cases} -1, & \text{if } s \notin S^T \\ 20, & \text{if } s \in S^T_{success} \\ 0, & \text{if } s \in S^T_{failure} \end{cases} \tag{1}$$

where $s \in S$, $a \in A$, $S^T \subset S$ is the set of terminal states, $S^T_{success} \subseteq S^T$, and $S^T_{failure} = S^T \setminus S^T_{success}$. A dialogue is considered successful if the retrieved item matched the user's preferences. $S^T_{success}$, therefore, contains all terminal states for which the dialogue is successful. While this reward function works well in controlled environments, in real-world applications it may not be possible to define $S^T_{success}$ and $S^T_{failure}$ as the true user's goal is unobservable. Moreover, $S^T_{success} \cap S^T_{failure}$ may not be empty, if the system partially meets the user's goal. Using surrogate methods to estimate metrics of dialogue quality, as proposed in this paper, has the potential to alleviate this shortcoming.

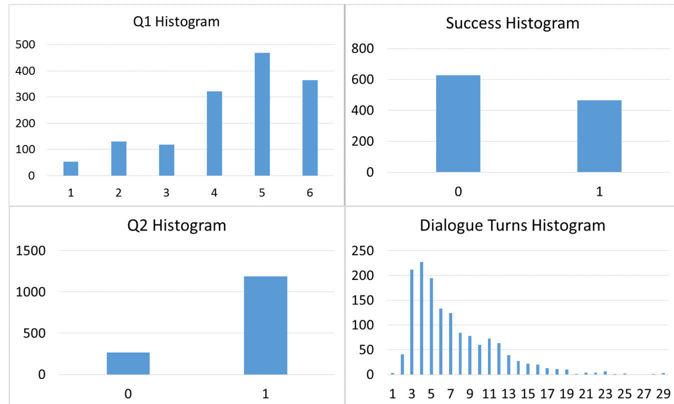

Fig. 1. Histograms showing the value distributions of the metrics of interest, in our dataset.

3. PREDICTING DIALOGUE SUCCESS, NATURALNESS, AND LENGTH

To train our models and predict the metrics of interest, we used a dataset of spoken dialogues between humans and a dialogue system, collected through Amazon Mechanical Turk (AMT).

3.1. Spoken Dialogue Data

The dataset consists of 1,456 dialogues (10,431 user utterances) between people and a statistical SDS. The interactions concerned finding appropriate Toshiba laptops or restaurants in Cambridge. Each person, therefore, was given a set of preferences, for example: *"You want a laptop for business use that is in the mid weight range. Make sure you get the size of its hard drive, and its dimensions"*. The person then interacted with the dialogue system until the item was retrieved or until the person decided to hung up. At the end of each dialogue, people were asked to provide feedback by answering the following two questions:

Q_1 *Did you find all the information you were looking for?*
Answer is on a 6-point Likert Scale.

Q_2 *The system understood me well.*
Answer is Yes or No.

In this work, we use Q_1 and Q_2 as indicators of user satisfaction [17, 21]. Therefore, a high mark in Q_1 and a 'Yes' in Q_2 indicate more satisfied users. Apart from these, two objective metrics were computed for each dialogue: dialogue success and number of (dialogue) turns. Dialogue success was determined by comparing the retrieved item against the set of preferences given to the user for the specific dialogue. A turn is defined as one system and one user utterance; a dialogue can therefore be thought of as a sequence of turns. Last, for each dialogue we have a complete log of the input and output to all components of the SSDS, from user's speech to system's speech (including partial ASR, SLU, beliefs, system actions, etc.) and we are thus able to extract acoustic and belief state features for each dialogue turn. Figure 1 shows the metrics' value distributions [1] in our dataset and Table 1 shows an example interaction.

[1]Note that there are some missing values in our dataset, e.g. success.

Acoustic Features	Belief State Features [20]
Mean of RMS	Top user dialogue act
St. dev. of RMS	Entropy of belief state slots
Mean of RMS derivative	System act
Mean pitch	Current turn
St. dev. of pitch	
Pitch range	
Diff. of mean pitch values in consecutive turns	
Mean of pitch derivative	

Table 2. Acoustic Features and Belief State Features.

3.2. Input Features

Acoustic Features. Our feature set is based on the Root Mean Square (RMS) and the pitch of the speech signal. More specifically, we used the following 8 acoustic features as input to our predictors: mean RMS (μ_{RMS}), mean pitch (μ_p), standard deviation of RMS (σ_{RMS}), standard deviation of pitch (σ_p), pitch range (r_p), difference of RMS in consecutive utterances (δ_{RMS}), difference of mean pitch values in consecutive utterances ($\delta_{\bar{p}}$), and the mean of the derivative of the pitch ($\mu_{\frac{dp}{dt}}$). RMS was computed using equation 2 and pitch information was extracted using the autocorrelation method (equation 3).

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x(t)^2} \tag{2}$$

where $\mathbf{x}(t) = \{x(1), ..., x(N)\}$ is the speech signal.

$$r_t(\tau) = \frac{1}{W}\sum_{j=t+1}^{t+W} x(j)x(j+\tau), \tag{3}$$

where t is the time the calculation is made, τ is the time lag, and W is the window size. The $r_t(\tau)$ function has a series of global maxima at zero, and then at all multiples of the period. The pitch period is determined by scanning this pattern, and is estimated by the location of the first global maximum with non-zero abscissa.

Belief State Features. We here briefly describe the features proposed in [20] that we used as a benchmark for our system. It should be noted here that [20] followed a different approach, utilising a simulated user and allowing their model access to turn-by-turn returns; this type of feedback is not possible in our case since we are using un-annotated spoken dialogue data. Regardless of this, [20] propose three feature sets containing information about the user's dialogue act, the system's dialogue act, current turn number and belief state information. The major difference among the three sets lies in the belief state information: a) F is defined as the full belief, b) F_{28} contains no belief state information, and c) F_{74} contains the entropy of each slot in the belief state. We selected F_{74} to compare against our approach, as it was the best performing set according to [20]. Table 2 summarizes the belief state feature set.

Acoustic and Belief State Features. Since the two feature sets described above essentially capture different kinds of information - relevant to *how the user spoke* (acoustic features) and *the perceived meaning of what they said* (belief state features) - we also explore how a joint feature set fares in predicting our metrics. An early fusion strategy, specifically feature-level fusion is adopted here. In other words, the fused feature vector is the concatenation of the acoustic features and the belief state features.

3.3. Metrics

As mentioned in the previous section, we use Q_1 (**subjective dialogue success**) and Q_2 (**interaction naturalness**) as constituents of user satisfaction. Other than these two metrics, we also predict the **objective dialogue success** and the **number of dialogue turns**. Since it was not feasible to ask users to provide ratings (i.e. answer questions Q_1 and Q_2) after every dialogue turn, we do not have a reliable way of telling "how responsible" each dialogue turn is for the final user rating. This means that we are not likely to achieve good accuracy in our predictions if we train our models on each utterance. To address this problem, we use the summary statistics of each feature over the course of a dialogue. Therefore, we calculated some summary statistics for each acoustic and belief state feature, specifically: mean, minimum and maximum value, standard deviation, skewness and kurtosis. Table 3 shows the accuracy of our predictors when using the acoustic features, the belief features or both sets of features. We deliberately selected mean over median for our summary statistics, as the median is less affected by sudden peaks, thus making it less informative in our case. This assumption was confirmed by running our models with median instead of mean and achieving lower prediction accuracies (not reported here).

4. EXPERIMENTS

We tried several classification and regression methods to predict the four metrics, and the best performing were binary SVM (with radial basis function or polynomial kernels), Gaussian Process Regressors (GPR, with squared exponential kernels), and Random Forests (RF). All our experiments were conducted with a 75-25 training-testing protocol, averaged over 10 repetitions. We scaled Q_1 and *Number of Turns* that take multiple values, into various scales in order to see at what level of granularity we can reliably make good predictions. For example, we scaled the *Number of Turns* from $\{1, ..., 29\}$ to $\{0, ..., 2\}$ which can be interpreted as "low", "medium" or "high" number of turns. Q_1 was scaled in a similar fashion. Specifically for the *Number of Turns*, we defined a binary metric that indicates when the dialogue is shorter than the mean length of successful dialogues:

$$DialogueLength(d) = \begin{cases} 1, & \text{if } NT_d \leq \mu_{NT_{D_s}} \\ 0, & \text{if } NT_d > \mu_{NTD_s} \end{cases} \tag{4}$$

where NT_d is the number of turns of dialogue $d \in D$, $D_s \subseteq D$ is the set of successful dialogues, and $\mu_{NT_D} = \frac{1}{|D|}\sum_d^D \{NT_d\}$.

4.1. Results

Table 3 summarises the results of our experiments, when using the summary statistics of the acoustic features extracted from users' utterances and using scaled versions of the metrics of interest. In the same table, the results for the belief state features and the fusion of acoustic and belief state features are demonstrated. We see that for binary metrics GPRs are the best performing models. GPR also performs best on all non-binary features, except when classifying the unscaled Q_1 responses, where RF perform best. Regarding the features sets, with the exception of predicting the number of turns scaled to $\{0, 1\}$, it is evident that belief state features alone cannot perform as well as the acoustic features or the joint acoustic and belief feature set. In fact, in some cases adding belief state information may not be helpful and can indeed hurt accuracy. This is a significant finding because it shows that acoustic features are rich in information that is useful for predicting constituents of user satisfaction.

Metric	Alg.	AF	BF	ABF
Q_1 $\{0,1\}$	SVM	0.691	0.696	0.677
	GPR	**0.746**	0.697	**0.746**
	RF	0.704	0.683	0.698
Q_1 $\{0-2\}$	GPR	**0.700**	0.580	0.658
	RF	0.582	0.577	0.574
Q_1 $\{0-6\}$	GPR	0.334	0.295	0.333
	RF	**0.400**	0.296	0.373
Q_2	SVM	0.864	0.816	0.817
	GPR	0.882	0.867	**0.903**
	RF	0.831	0.835	0.818
Success	SVM	0.579	0.571	0.580
	GPR	**0.775**	0.766	0.774
	RF	0.586	0.576	0.621
DialogueLength	SVM	0.704	0.762	0.714
	GPR	**0.951**	0.795	0.671
	RF	0.880	0.837	0.926
Turns $\{0,1\}$	SVM	0.950	0.961	0.947
	GPR	0.947	0.948	0.979
	RF	0.945	**0.982**	0.952
Turns $\{0-2\}$	GPR	**0.972**	0.880	0.964
	RF	0.846	0.895	0.879
Turns $\{0-3\}$	GPR	0.922	0.741	**0.945**
	RF	0.792	0.845	0.837
Turns $\{0-29\}$	GPR	**0.716**	0.193	0.389
	RF	0.492	0.619	0.619

Table 3. Classification results on the various metrics, when using the summary statistics of the acoustic features (AF), the belief state features (BF) or the all of them (ABF).

Another important observation is that even though the acoustic features do not have access to the current turn number, they perform better than the belief state features which do include such information when predicting the *DialogueLength* metric. AF or ABF features also perform better in predicting the number of turns [2], compared to BF. This could be because AF capture changes in the speakers' signal when the dialogue is longer than usual (e.g. speakers may be getting annoyed or impatient).

A general trend, therefore, is that AF lead to better accuracy than what can be achieved with BF alone. As further evidence to support this, we examine the confusion matrices, where due to space limitations we focus on the *DialogueLength* metric. The confusion matrix when AF are used can be seen in Table 4, for the BF can be seen in Table 5, whereas the feature-level fusion results are demonstrated in Table 6. Since we adapted a holdout validation with stratification, 10 individual confusion matrices are produced. However, here we report the mean value followed by the standard deviation for each of the individual elements of the confusion matrix, e.g. the mean of the 10 values of the dialogues that had a length less than $\mu_{NT_{D_s}}$ and the algorithm predicted so. It can be deducted that the correctly classified instances when the AF are utilised (Table 4) are more than when the BF are used (Table 5). However, combining the two types of features (ABF) leads to more correctly classified instances, as shown in Table 6. We also observe that the standard deviation is low, especially for the case of the correctly classified instances, thus proving the stability and robustness of the proposed method.

[2]Again with the exception of number of turns scaled to $\{0, 1\}$.

True DL \ Predicted DL	$NT_d \leq \mu_{NT_{D_s}}$	$NT_d > \mu_{NT_{D_s}}$
$NT_d \leq \mu_{NT_{D_s}}$	199.3 (1.3)	3.3 (1.2)
$NT_d > \mu_{NT_{D_s}}$	43.3 (6.3)	118.1 (6.4)

Table 4. Confusion matrix for the RF on Dialogue Length (DL) when using the AF set.

True DL \ Predicted DL	$NT_d \leq \mu_{NT_{D_s}}$	$NT_d > \mu_{NT_{D_s}}$
$NT_d \leq \mu_{NT_{D_s}}$	171.0 (3.7)	31.5 (3.8)
$NT_d > \mu_{NT_{D_s}}$	29.3 (4.9)	132.2 (5.1)

Table 5. Confusion matrix for the RF on Dialogue Length (DL) when using the BF set.

True DL \ Predicted DL	$NT_d \leq \mu_{NT_{D_s}}$	$NT_d > \mu_{NT_{D_s}}$
$NT_d \leq \mu_{NT_{D_s}}$	196.9 (2.5)	5.7 (2.2)
$NT_d > \mu_{NT_{D_s}}$	19.0 (4.0)	142.4 (3.8)

Table 6. Confusion matrix for the RF on Dialogue Length (DL) when using the ABF set.

Contrary to [19, 20], who use simulated data to train deep neural networks (DNN), we use acoustic data from human users which are richer in information but also significantly less in terms of number of dialogues. Therefore, we could not apply DNN methods and this is why our work cannot be directly compared to [19, 20]. In both of these works, the authors apply their pre-trained RNN models to train dialogue policies on-line but do not learn the RNN models themselves from real user data. Another difference lies in the nature of the data which restricts the training methods. Specifically, we did not have access to per-turn return for the two user-provided metrics, Q_1 and Q_2; we therefore used features extracted for each dialogue. The output of our predictors, however, can be used to train end to end SSDS or specific components such as dialogue policies.

5. CONCLUSION

We proposed an innovative method to estimate metrics related to user satisfaction and dialogue quality when interacting with SSDS, that uses simple acoustic features and achieves better performance than when using belief state features as is the current state of the art. The proposed method is the first, to the best of the authors' knowledge, to predict such metrics by exploiting audio features. The methodology combines signal processing for feature extraction, namely RMS and pitch related features with supervised classification, specifically SVMs, GPRs and RFs. Results indicate that using audio features enhances the classifiers' performance either when used alone or in conjunction with belief state features.

In the future, we will incorporate the proposed acoustic features in the dialogue state and let policy learning (whether RL or DNN) solve the credit assignment problem while optimising for dialogue success. Moreover, we will explore linguistic features that could be strong predictors of our metrics. This richer dialogue state will be used to guide not only the system's output at *system act* level, but provide information to Text To Speech and Language Generation as well, in an effort to handle dialogues other than information-seeking.

6. REFERENCES

[1] Blaise Thomson, *Statistical methods for spoken dialogue management*, Springer Science & Business Media, 2013.

[2] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

[3] Paul A Crook, Simon Keizer, Zhuoran Wang, Wenshuo Tang, and Oliver Lemon, "Real user evaluation of a pomdp spoken dialogue system using automatic belief compression," *Computer Speech & Language*, vol. 28, no. 4, pp. 873–887, 2014.

[4] Kallirroi Georgila, Claire Nelson, and David R Traum, "Single-agent vs. multi-agent techniques for concurrent reinforcement learning of negotiation dialogue policies.," in *ACL (1)*, 2014, pp. 500–510.

[5] Heriberto Cuayahuitl, Simon Keizer, and Oliver Lemon, "Strategic dialogue management via deep reinforcement learning," 12 2015.

[6] Merwan Barlier, Julien Perolat, Romain Laroche, and Olivier Pietquin, "Human-machine dialogue as a stochastic game," in *16th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2015)*, 2015.

[7] Yun-Nung Chen, Ming Sun, Alexander I Rudnicky, and Anatole Gershman, "Unsupervised user intent modeling by feature-enriched matrix factorization," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6150–6154.

[8] Tiancheng Zhao and Maxine Eskenazi, "Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning," *to appear in SIGDial*, 2016.

[9] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve J. Young, "A network-based end-to-end trainable task-oriented dialogue system," *CoRR*, vol. abs/1604.04562, 2016.

[10] Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *AAAI*, 2016.

[11] Raul Fernandez and Rosalind W. Picard, "Dialog act classification from prosodic features using support vector machines," in *Proceedings of Speech Prosody*, 2002, SP '02, pp. 291–294.

[12] Alessandra Cervone, Catherine Lai, Silvia Pareti, and Peter Bell, "Towards automatic detection of reported speech in dialogue using prosodic cues," in *INTERSPEECH '15*, 2015, pp. 3061–3065.

[13] Joseph Tepperman, David Traum, and Shrikanth Narayanan, ""yeah right": Sarcasm recognition for spoken dialogue systems," in *Nineth Annual Conference of the International Speech Communication Association*, 2006, INTERSPEECH '06, pp. 1838–1841.

[14] Jose Lopes, Giampiero Salvi, Gabriel Skantze, Alberto Abad, Joakim Gustafson, Fernando Batista, Raveesh Meena, and Isabel Trancoso, "Detecting repetitions in spoken dialogue systems using phonetic distances," in *INTERSPEECH '15*, 2015, pp. 1805–1809.

[15] Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman, "Learning to predict problematic situations in a spoken dialogue system: experiments with how may i help you?," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics, 2000, pp. 210–217.

[16] Narichika Nomoto, Hirokazu Masataki, Osamu Yoshioka, and Satoshi Takahashi, "Detection of anger emotion in dialog speech using prosody feature and temporal relation of utterances," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[17] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella, "Paradise: A framework for evaluating spoken dialogue agents," in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997, pp. 271–280.

[18] Layla El Asri, Romain Laroche, and Olivier Pietquin, "Task completion transfer learning for reward inference," *Proc of MLIS*, 2014.

[19] Pei-Hao Su, David Vandyke, Milica Gasic, Dongho Kim, Nikola Mrksic, Tsung-Hsien Wen, and Steve Young, "Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems," *Interspeech*, 2015.

[20] David Vandyke, Pei-Hao Su, Milica Gasic, Nikola Mrksic, Tsung-Hsien Wen, and Steve Young, "Multi-domain dialogue success classifiers for policy training," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 763–770.

[21] Marilyn A Walker, Rebecca Passonneau, and Julie E Boland, "Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2001, pp. 515–522.

[22] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1, pp. 99–134, 1998.

[23] Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu, "The hidden information state model: A practical framework for pomdp-based spoken dialogue management," *Computer Speech & Language*, vol. 24, no. 2, pp. 150–174, 2010.