Enriching text-to-speech synthesis using automatic dialog act tags

Vivek Kumar Rangarajan Sridhar, Ann Syrdal, Alistair Conkie, Srinivas Bangalore

AT&T Labs - Research
180 Park Avenue, Florham Park, NJ 07932
vkumar,syrdal,adc,srini@research.att.com

Abstract

We present an approach for enriching dialog based text-to-speech (TTS) synthesis systems by explicitly controlling the expressiveness through the use of dialog act tags. The dialog act tags in our framework are automatically obtained by training a maximum entropy classifier on the Switchboard-DAMSL data set, unrelated to the TTS database. We compare the voice quality produced by exploiting automatic dialog act tags with that using human annotations of dialog acts, and with two forms of reference databases. Even though the inventory of tags is different for the automatic tagger and human annotation, exploiting either form of dialog markup generates better voice quality in comparison with the reference voices in subjective evaluation.

Index Terms: dialog acts (speech acts), unit selection speech synthesis, prosody

10.21437/Interspeech.2011-119

1. Introduction

In both human-to-human and human-computer speech communication, identifying whether an utterance is a statement, question, greeting, etc., is integral to producing, sustaining and understanding natural dialogs. Hence, it is important that text-to-speech synthesis systems in a dialog setting be able to synthesize utterances with appropriate dialog context. Among various approaches to text-to-speech synthesis, unit selection synthesis [1] has been the most popular in recent times due to it's ability to produce natural sounding speech. The underlying assumption is that the unit database is rich in units that cover a varying degree of expressiveness and hence the synthesized utterance will be rendered with appropriate prosody. In a dialog setting, this is akin to constructing a unit database with utterances that cover all categories of dialog types. While this notion of controlling prosody in conventional unit selection synthesis systems is implicit, explicit control over the expressiveness and prosody can be obtained by either using specialized databases or categorical mark up tags that constrain the unit selection search.

Dialog act (speech act) tags are labels used to represent such surface level communicative acts in a conversation or dialog [2]. Exploiting such information based on pragmatic categories has been shown to improve the quality of text-to-speech synthesis output [3]. For instance, assume that the unit database is recorded using a representative set of dialog types such as *statements*, *yes-no questions*, *greetings*, etc. An input text "Hi! it's Annie. How may I help you?" may be synthesized using a sequence of units from *statement* utterance type. Furthermore, the pitch range of the synthesized utterance may be based on global statistics of the unit database. However, explicitly adding dialog act tags can result in "Hi! it's Annie. <*conventional-opening*> How may I help you? <*wh-question*>" being synthesized with appropriate set of units belonging to a particular type of dialog act and the pitch range dependent on the dialog type.

In this paper, we focus on enriching unit selection synthesis in a dialog scenario by incorporating explicit knowledge about dialog context through dialog act tags. Our objective is two-fold. First, we circumvent the bottleneck of manual annotation of TTS prompts [3, 4] by using an automatic dialog act tagger. Second, we investigate the utility of a generalized tag set [5] for representing dialog categories in comparison with one designed specifically with respect to a TTS database [3]. The automatic dialog act tagger in this work is trained on the Switchboard-DAMSL corpus [5]. The dialog act tags are used within the unit selection system in two ways: (i) The dialog act tag is used as a feature during the unit selection search, i.e., the synthesized unit sequences are dependent on the dialog act tags; (ii) we control the prosody of the synthesized utterances by modifying the pitch range based on the dialog act specific ranges observed in labeled TTS database.

We compare the voice quality produced with the automatic dialog act tags with that of two baseline voices without any markup, as well as a voice constructed from a human annotated dialog act database [3, 4]. Subjective evaluation results demonstrate that the voice quality of the TTS system using either the automatically tagged or human annotated database is better than the reference voices without any markup. The proposed scheme for utilizing automatic dialog act tags in TTS is scalable and is a simple way of providing expressiveness in dialog applications. We present details about the automatic dialog act tagging framework: models, data and tagging accura-

cies in Section 2 followed by the description of the unit selection database in Section 3. Section 4 describes the use of dialog act tags in in AT&T Natural Voices speech synthesis system [6]. We present experimental results in Section 5 and conclude in Section 6 along with some directions for future work.

2. Automatic Dialog Act Tagging

Automatic dialog act tagging pertains to classifying an utterance (either text or both speech and text) into a one of many semantic categories that signify the surface level communicative act. Typically, supervised learning procedures are used to learn the mapping from the utterance (represented through a set of features) into the dialog act category. In this work, we use a maximum entropy tagger to perform this classification.

2.1. Maximum Entropy Model

We model the prediction problem as a classification task: given a sequence of utterances u_i in a dialog $U = u_1, u_2, \cdots, u_n$ and a dialog act vocabulary ($d_i \epsilon \mathcal{D}, |\mathcal{D}| = K$), we need to predict the best dialog act sequence $D^* = d_1, d_2, \cdots, d_n$.

$$D^* = \arg\max_{d_1,\cdots,d_n} P(d_1, \cdots, d_n | u_1, \cdots, u_n) \quad (1)$$

We approximate the sequence level global classification problem, using conditional independence assumptions, to a product of local classification problems as shown in Eq.(2). The classifier is then used to assign to each word a dialog act label conditioned on a vector of local contextual feature vectors comprising the lexical, syntactic and acoustic information.

$$D^* \approx \arg\max_{D} \prod_{i=1}^{n} P(d_i | \mathbf{\Phi}(u_{i-k}, \cdots, u_{i+l})) \quad (2)$$

$$= \arg\max_{D} \prod_{i=1}^{n} P(d_i | \mathbf{\Phi}(W_{i-k}, \cdots, W_{i+l}, S_{i-k}, \cdots, S_{i+l})) \quad (3)$$

where W_i is the word sequence and S_i is the syntactic feature sequence belonging to utterances u_i. The variables l and k denote the right and left context respectively.

To estimate the conditional distribution $P(d|\mathbf{\Phi})$ we use the general technique of choosing the maximum entropy (maxent) distribution that estimates the average of each feature over the training data. We use the machine learning toolkit LLAMA [7] to estimate the conditional distribution using maxent.

2.2. Training Data for Dialog Act Tagger

We use the Switchboard-DAMSL (SWBD-DAMSL) [5] corpus as the training data to train our dialog act tagger. The SWBD-DAMSL corpus consists of 1155 dialogs and 218,898 utterances from the Switchboard corpus of telephone conversations, tagged with discourse labels from a shallow discourse tag set. It contains 42 dialog act tags that distinguish mutually exclusive utterance types [5]. The interlabeler agreement for this 42-label tag set is 84% ($\kappa = 0.80$), with the labeling performed at the utterance level. In order to benchmark the accuracy of the tagger described in Section 2.1, we used a set of 173 dialogs, selected at random for testing. The test set consisted of 29869 discourse segments.

2.3. Tagging Results

The lexical features used in our modeling framework are trigrams of words in a given utterance. We tag the utterances with part-of-speech tags using the AT&T POS tagger. The POS inventory is the same as the Penn treebank which includes 47 POS tags. In addition to the POS tags, we also annotate the utterance with Supertags [8]. The syntactic features comprise the trigrams of POS and Supertags.

For more detailed information about the dialog act tagger, the reader is referred to [9]. However, the results presented in this paper are different from those presented in [9]. The performance improved significantly as we used punctuation as feature in the classification framework. In previous work we omitted punctuation in the training data as our test domain was the output of automatic speech recognition that typically does not generate punctuation.

Cues used	**Accuracy**
Chance (majority tag)	39.9
Lexical	74.9
Lexical+Syntactic	76.0

Table 1: Dialog act tagging accuracies (in %) for lexical and syntactic cues with the maximum entropy model.

Results of DA tagging using lexical and syntactic features from reference transcripts are presented in Table 1. Analysis of the errors made by the classifier demonstrated that majority of the errors were associated with low frequency tags. The high frequency dialog act tags such as *statement-non-opinion*, *wh-question*, *yes-no question*, *acknowledgment*, etc. exhibited higher accuracies ($\approx$ 82-85%).

3. Unit Selection Database

The TTS experiments in this work used a 12 hour speech corpus recorded from an adult female speaker who was a native speaker of American English. The corpus consisted of text typically spoken in human-computer dialog applications. The texts included dialogs transcribed from customer-live agent interactions, simulated dialogs

based on such interactions, prompts for interactive services, and information requested from automated interactive services, such as names, addresses, flight information, digit strings such as used for telephone, account, or credit card numbers.

The dialog act of every utterance in the 12 hour corpus was annotated manually using a set of 20 tags specifically designed for the data set [4]. The utterances were also classified into dialog act categories using the automatic dialog act tagger presented in Section 2. A total of 24 tags out of the 42 tag vocabulary of the tagger was used in labeling the TTS database.

4. Integrating dialog act tags in TTS

The dialog act tags from both the automatic tagger and human annotation were used in two ways for building the unit selection voice. First, the written text corresponding to the acoustic inventory was processed to provide labels that were aligned with the acoustic units. The only distinguishing factor between automatic and human annotated tags was in terms of the tag set used. The set of dialog act labels was made available to the unit selection system so that any text to be synthesized with mark-up could be compared with database units in terms of conversional features (e.g., f0, duration) and additionally in terms of the dialog act feature. A weight was provided so that the importance of the feature could be adjusted. Second, the prosodic variables (pitch range parameters) specified for the voice were different for each dialog act tag. These values were inferred from the labeled unit selection database.

5. Experiment

Web-based listening testing was conducted to test the utility of the automatic dialog tag system for improving the TTS quality in the context of human-computer dialogs. Two distinct tests were performed in the evaluation. One investigated the use of dialog act tags for two short dialogs while the other examined using tags in a more general context. The details of the stimuli and design of the experiment are presented in the following sections.

5.1. Stimuli

The automated agent portion of two simulated dialogs, one a travel reservation scenario and the other a restaurant-booking scenario, and of a mixed set of unrelated utterances, were synthesized using four different TTS systems: (1) Standard TTS (STD) used the standard AT&T Research unit selection TTS system, (2) Manual Speech act TTS (MSA) used the human annotated speech acts, (3) Speech act TTS using automatic tags (ASA) and (4) a system built with the same acoustic inventory as (2) and (3), but lacking any speech act information or adaptations (NSA). All four systems were built from the same speaker, although the recorded material included in the first, standard, inventory differed in both size and constituent material from that of the other three systems. The standard TTS inventory contained approximately 6 hours of speech; the recorded material was primarily reading of factual material, but it also included some interactive dialog material.

From each of the above four systems, twenty two utterances (two sets of seven, representing agent turns in a dialog and one set of eight, representing a mixed set) were generated and used as listening test stimuli. The input text was standard text for STD and NSA. The text was annotated manually for MSA and automatically for ASA.

5.2. Design

The Web-based listening tests were administered in two ways: Web interface hosted on a standalone server and Amazon Mechanical Turk. The listening tests were divided into three parts.

(a) Travel dialog (seven utterances)

(b) Restaurant search dialog (seven utterances)

(c) General utterances (eight utterances).

Each part consisted of paired comparisons in which listeners rated their A/B preference on a -2 (strongly prefer A) to +2 (strongly prefer B) scale, where 0 indicated no preference. Sequential order of the seven dialog turns was preserved for (a) and (b) but utterance order was randomized for the eight pairs in (c). Stimuli were presented using a Latin square design (for the 4 TTS systems) to ensure as far as possible a balanced set of paired comparisons, including order of presentation. The three tests over the 4x4 Latin Square design were posted as individual human intelligence task (HIT) on Mechanical Turk. In all three tests, listeners also indicated whether or not English was their native language, and whether they listened using headphones or speakers.

5.3. Listeners

Preference results are based on 5121 ratings from 148 unique listeners for dialog tests (a and b), and on 3043 ratings from 111 listeners in mixed test c. Many of the participated in more than one test page. Of the initial 182 test participants in dialog tests and 136 participants in test c, 34 were judged unreliable and their data were eliminated from analysis in dialog tests, and 25 were eliminated for test c. Listeners were judged unreliable if (a) their ratings never varied (e.g. all ratings were "-2" or all "+2"), or (b) if they strongly preferred one of two paired identical stimuli over 40% of the time, or (c) if they took multiple tests but were inconsistent in reporting their native language status. Of the ratings from acceptably reliable listeners, 3990(78%) in dialog tests, and 2374(78%) in test

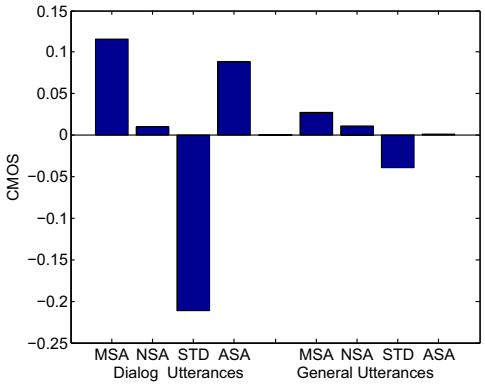

Figure 1: *CMOS for four TTS systems*

c reported using headphones or earphones; the remaining 22% listened by loudspeaker. 3762(73%) identified themselves as native speakers of English in dialog tests, and 2275(75%) in test c; the remaining listeners identified themselves as non-native.

5.4. Results

Overall quality as measured by Comparison Mean Opinion Scores (CMOS) is shown in Figure 1. The potential CMOS range is -2 to 2. $CMOS = (2N_{+2} + N_{+1} + 0N_0 - N_{-1} - 2N_{-2})/N_{tot}$, where N_{-2} is the number of 'strongly prefer A' (-2) responses, N_1 is the number of 'prefer B' (+1) responses etc, and N_{tot} is the total number of responses. Note that in this test, CMOS was calculated from ratings of each of the two members of an A/B pair (e.g. for a rating of 2, A would score -2 and B would score 2).

Comparative Mean Opinion Scores (CMOS) are shown in Figure 1, with dialog tests (a) and (b) combined, and test (c) shown separately. The results show a much greater range of variation for the dialog test stimuli in comparison with the general utterance test set. Scores on the dialog tests were highest for the two speech act systems (MSA and ASA). The STD reference system was strongly disfavored by listeners, as indicated by its strongly negative CMOS value. The NSA system's CMOS score was just above the neutral score (0). On the other hand, the general utterance test set exhibits much smaller range in CMOS scores. Hence, using dialog act tags in the appropriate context can enhance the quality of the synthesis as demonstrated by the CMOS scores.

6. Summary and Conclusions

We presented an approach for incorporating expressiveness in dialog-based TTS by explicitly using dialog act tags generated by an automatic tagger using the Switchboard-DAMSL tag set. The TTS database was then constructed using the tagged data. We compared our approach with three different TTS systems, one constructed using a manually tagged database (with customized tags) while the other two were reference systems that differed in the amount of training data. Subjective evaluation showed that the automatic system achieved similar performance in comparison with the manually tagged system and notably better than both the reference systems in the dialog domain. The automatic framework provides a means to scale to larger unit selection databases and is suitable as a preprocessing module for enriching general purpose text-to-speech synthesizers. We plan to explore in more detail the effects of individual speech acts on the voice quality as part of future work.

7. References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large database," in *Proceedings of ICASSP*, 1996.

[2] J. L. Austin, How to do Things with Words. Clarendon Press, Oxford, 1962.

[3] A. K. Syrdal, A. Conkie, Y.-J. Kim, and M. Beutnagel, "Speech acts and dialog TTS," in *Proceedings of Speech Synthesis Workshop*, 2010.

[4] A. K. Syrdal and Y.-J. Kim, "Dialog speech acts and prosody: Considerations for TTS," in *Proceedings of Speech Prosody*, 2008.

[5] D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, S. Stolcke, P. Taylor, and C. Van Ess-Dykema, "Switchboard discourse language modeling project report," Center for Speech and Language Processing, Johns Hopkins University, Technical Report Research Note 30, 1998.

[6] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Proceedings of Joint Meeting of ASA, EAA and DEGA*, 1999.

[7] P. Haffner, "Scaling large margin classifiers for spoken language understanding," *Speech Communication*, vol. 48, pp. 239–261, 2006.

[8] S. Bangalore and A. Joshi, "Supertagging: An approach to almost parsing," *Computational Linguistics*, vol. 25, no. 2, 1999.

[9] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging," *Comput. Speech Lang.*, vol. 23, pp. 407–422, October 2009.