DIALOGUE CONTEXT SENSITIVE HMM-BASED SPEECH SYNTHESIS

Pirros Tsiakoulis, Catherine Breslin, Milica Gašić, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Steve Young

University of Cambridge, Engineering Department, Cambridge, UK
pt344@cam.ac.uk

ABSTRACT

The focus of this work is speech synthesis tailored to the needs of spoken dialogue systems. More specifically, the framework of HMM-based speech synthesis is utilized to train an emphatic voice that also considers dialogue context for decision tree state clustering. To achieve this, we designed and recorded a speech corpus comprising system prompts from human-computer interaction, as well as additional prompts for slot-level emphasis. This corpus, combined with a general purpose text-to-speech one, was used to train voices using a) baseline context features, b) additional emphasis features, and c) additional dialogue context features. Both emphasis and dialogue context features are extracted from the dialogue act semantic representation. The voices were evaluated in pairs for dialogue appropriateness using a preference listening test. The results show that the emphatic voice is preferred to the baseline when emphasis markup is present, while the dialogue context-sensitive voice is preferred to the plain emphatic one when no emphasis markup is present and preferable to the baseline in both cases. This demonstrates that including dialogue context features for decision tree state clustering significantly improves the quality of the synthetic voice for dialogue.

Index Terms: HMM-based speech synthesis, emphatic speech synthesis, dialogue context-sensitive speech synthesis

1. INTRODUCTION

Speech has gained significant ground as a human-machine interface, enabling Spoken Dialogue Systems (SDS) for a variety of applications [1]. Such systems often employ a general purpose synthetic voice with neutral characteristics. Recent effort has focused on making the discourse more natural, incorporating spontaneous responses, backchannel and fillers, as well as incremental processing [2, 3, 4, 5, 6, 7]. This pinpoints the need for expressive speech synthesis that is aware of the discourse context [8]. The generated system prompts need to be concise and convey more information via synthesis, dialogue context-sensitive speech synthesis [9]. The Text-to-Speech (TTS) component of a spoken dialogue system is typically preceded by the Natural Language Generation (NLG) component. The NLG component translates the intended dialogue action from a high-level semantic representation into text. This facilitates richer generation; in addition to plain text, the NLG component can also produce expressive annotations [10, 11]. However, expert knowledge and effort is required to design and implement both the NLG and TTS components.

This paper investigates the potential of an expressive TTS component targeting the needs of a spoken dialogue system without the need of any complex annotation scheme. Instead, the existing dialogue act semantic representation is used as an additional contextual factor for decision tree state clustering in HMM-based speech synthesis. This work mainly considers emphasis and style as the target aspects of expressive speech for dialogue. Emphasis provides a way of highlighting the focus of the utterance and naturally signalling what the user should pay attention to. Style, on the other hand, which can be manifested in various ways, e.g. speaking rate, pitch variations, etc., can be used to convey more subtle information to the user. For example, the speaking rate may be reduced (in conjunction with emphasis) when giving new information to the user.

To this end, a new speech corpus was collected for expressive speech generation within the dialogue domain. The corpus includes system-user pairs of interaction prompts from previously collected dialogues, as well as individual prompts designed specifically for emphasis patterns. A professional speaker was instructed to act as the dialogue system operator and convey information to the user using contextually appropriate speech. The collected speech corpus was used in addition to a general purpose text-to-speech corpus to build: a) a voice using baseline context features, b) an emphatic voice by including slot-level emphasis context features, and c) a dialogue context-sensitive emphatic voice by including contextual factors extracted from the intended dialogue act semantic representation. A live user trial was ineffective in assessing the utility of the voices in a dialogue system, hence a preference listening test was designed. A dialogue was presented to the user where each system turn had a pair of alternative synthetic prompts. The user was asked to choose the most appropriate system response or indicate no preference. The results show that a) the emphatic voice is preferred to the baseline when emphasis markup is present, b) the context-sensitive voice is preferred to the plain emphatic one when no emphasis markup is present, and c) the context-sensitive voice preferred to the baseline in both scenarios. The preference towards the dialogue context-sensitive voice is consistent across different dialogue act types regardless of the emphasis status. This demonstrates that dialogue context features can be used in conjunction with emphasis features to improve the quality of synthetic speech for dialogue.

1.1. Related Work

The idea of semantic input to the speech synthesizer was originally introduced by Young and Fallside using the term Speech Synthesis from Concept [12]. The term Concept-To-Speech (CTS) later prevailed to describe methods that combine joint NLG and TTS functionality. One approach to CTS involves an annotation schema which is applied to the generated text, and affects the prosody of the rendered speech [10]. A similar technique applies prosodic annotations to a template-slot based generation system [11]. Another approach is to jointly optimize text and prosody generation in the framework of unit selection TTS [13, 14]. Others have focused on prosody models for CTS, which are driven from semantic input as
well as linguistic input [15, 16, 17]. Our approach is not strictly a
CTS one, since it does not require any complex annotation schema,
or strong coupling between NLG and TTS. Instead, the semantic rep-
resentation of the dialogue acts is used to extract context features for
decision tree state clustering in HMM-based speech synthesis.

There is a considerable amount of ongoing research on HMM
based statistical speech synthesis (HTS) [18], which has led to sig-
nificant improvement in the quality of the synthetic speech [19].
HTS uses decision trees to cluster and model the acoustic-prosodic
space. The decision trees are built in a data-driven manner using lin-
guistic information extracted from text. Any paralinguistic or non-
linguistic information can be used as long as it can be predicted from
text or input otherwise. In this paper, the HTS framework is utilized
to investigate the use of dialogue and emphasis information that is
directly extracted from the dialogue act representation.

Several efforts for modeling emphasis have been proposed in
the framework of HMM-based speech synthesis. In most cases, a
data-driven approach is followed, either by detecting/annotating em-
phased words in existing corpora [20, 21] or by collecting speech
corpora specifically designed for emphasis modeling [22]. Empha-
sis context features are then used in the decision tree state clustering
stage. More elaborate techniques have also been proposed that can
tackle data sparsity issues when the emphasis data is limited, such
as factorized decision trees [21, 23], hierarchical modeling [24],
and phrase level modeling [25]. Adaptation techniques have also
been proposed for different aspects of expressive speech synthesis
[26, 27]. The goal of this work is not to propose a new technique,
but rather explore existing ones in the context of a dialogue system.

2. EXPRESSIVE DIALOGUE CORPUS

The restaurant domain was selected as the primary application do-
main, mainly because of data availability. Emphasis and style were
selected as the primary expressive patterns to be covered. The scripts
to be recorded were annotated to indicate words that should be em-
phazized. The expressive style, on the other hand, is neither strictly
defined, nor is an annotation scheme available. Specifically for the
dialogue domain, the expressive style should reflect the current dia-
logue state, e.g. the confidence level. These phenomena were mod-
eled implicitly by including whole dialogues in the corpus, in order
to utilize dialogue context features during voice training.

2.1. Existing Dialogue Corpora

The initial source data consisted of previously collected dialogues
using the Cambridge spoken dialogue system. This data is summa-
rized in Table 1. The TownInfo domain includes restaurant, hotel and
bar information for a hand-crafted information database [28], while the
TopTable domain contains restaurants provided by an online ser-
vice provider [29]. An initial investigation into using a subset of this
dialogue data showed that it was not rich enough for the purpose at
hand. It had limited diversity in terms of the system prompts as well
as the venue names. Note that counts of unique prompts in Table 1
were calculated including the actual slot values used, e.g. names and
numbers. Therefore the prompts were preprocessed and enriched.

<table>
<thead>
<tr>
<th>Domain</th>
<th># Dialogues</th>
<th># Turns</th>
<th># Unique Prompts</th>
</tr>
</thead>
<tbody>
<tr>
<td>TownInfo</td>
<td>1422</td>
<td>13992</td>
<td>3346</td>
</tr>
<tr>
<td>TopTable</td>
<td>2166</td>
<td>28846</td>
<td>2284</td>
</tr>
<tr>
<td>Total</td>
<td>3588</td>
<td>42838</td>
<td>5614</td>
</tr>
</tbody>
</table>

Table 1. Source dialogue data for the corpus design.

2.2. Prompt Processing and Corpus Selection

In order to add some variety into the design of the final corpus, the
extracted prompts were enriched semi-automatically. The transforma-
tion procedure included the following steps:

- Extract pairs of dialogue act and corresponding system
  prompt, e.g.
  `inform(name="la tasca", postcode="CB3 0AD")`
  `The postcode of la tasca is CB3 0ad`

- Replace slot values with slot class names, e.g.
  `inform(name=NAME, postcode=POSTCODE)`
  `The postcode of NAME is POSTCODE`

- Provide alternatives by rephrasing the prompt, e.g.
  `The postcode of NAME is POSTCODE; Its postcode is POSTCODE; ...`

- Select a list of dialogues that maximize the coverage of the
  extended list of prompts. A simple greedy algorithm was used
  for this task. At each step, the algorithm added the dialogue
  which included the most unseen prompts to the list of the se-
  lected dialogues. This is similar to the standard approaches to
corpus design that operate at the word level [30].

- The slot class names were replaced with slot values. If a di-
  alogue involved a venue that had already been spoken about
  in a previously selected dialogue, the venue was randomly re-
  placed with another venue to avoid many repetitions of the
  same venue name. For other slots, such as count, phone num-
  ber, and postcode, a random list was generated.

- Some artificial turns were added to include free text descrip-
tions that were available for some of the venues.

The summary of the selected dialogues is shown in Table 2. In
addition, an extra set of prompts was selected for the emphasis task.
Each prompt was recorded multiple times, each time with a different
slot emphasized. An example is given below

Are you looking for a Portuguese restaurant in Barnwell ?
Are you looking for a Portuguese restaurant in Barnwell ?
Are you looking for a Portuguese restaurant in Barnwell ?
Are you looking for a Portuguese restaurant in Barnwell ?

2.3. Emphasis Assignment

The dialogue corpus was annotated using emphasis tags at the slot
level. More specifically, for every dialogue the first encounter of
each slot value was annotated with an emphasis tag. This assumes
that the system should emphasize every new bit of information that
it presents to the user1. The technique was also implemented and
integrated into the dialogue system which was used for evaluation.
An example dialogue is shown below where emphasized words are
marked in bold-face.

<table>
<thead>
<tr>
<th>Domain</th>
<th># Dialogues</th>
<th># Turns</th>
<th># Unique Prompts</th>
</tr>
</thead>
<tbody>
<tr>
<td>TownInfo</td>
<td>86</td>
<td>1089</td>
<td>407</td>
</tr>
<tr>
<td>TopTable</td>
<td>131</td>
<td>1351</td>
<td>1018</td>
</tr>
<tr>
<td>Total</td>
<td>217</td>
<td>2440</td>
<td>1425</td>
</tr>
</tbody>
</table>

Table 2. Dialogue data included in the final corpus.

1More elaborate methods can be applied at run-time by the NLG compo-
  nent to assign slot/word emphasis tags.
3. EXPERIMENTS

3.1. Emphasis and Dialogue Context Features

All voices were trained on the same dataset including the original RJS corpus and the new expressive dialogue corpus. The training setup was also kept the same using a modified version of HTS that incorporates continuous F0 modeling [32]. The following stream configuration was used: 25 Mel-Cepstral coefficients, log F0, five band aperiodic energy components, and voicing condition [32].

Three voices were trained using: a) baseline context features, b) additional emphasis features (6 questions [21]), and c) additional dialogue context features (including emphasis). Dialogue acts take the form \( \text{dact}(a_1[=0], \ldots, a_N[=0]) \), where \( \text{dact} \) is the dialogue act type, \( \{a_i, v_i\} \) is the \( i \)-th slot-value pair, and \( N \) is the number of slots, e.g. \( \text{inform}(\text{food}=\text{Chinese}), \text{request}(\text{area}) \). The dialogue context features are: the dialogue act type (17 additional questions), the type of food \( \text{food} \) (21), and the type of cuisine \( \text{cuisine} \) (21). The user had to interact with the system to get a venue matching the given constraints, and then to ask for the required information about that restaurant. More complex dialogues would occur if there were no matching venues, in which case the user could relax one of the given constraints. At the end of the dialogue, the user was asked to judge the dialogue for: a) task completion success (Yes or No), b) perceived comprehension on a five-point Likert scale from strongly disagree to strongly agree (if the system understood the user), c) overall impression of the quality of the system’s voice, d) emphasis assignment, and e) intonation. The latter three questions, which are relevant for TTS, were rated on a continuous scale (0-60) for Mean Opinion Score (MOS) [19].

Three systems were tested having identical configurations, except the synthetic voice used by the TTS component. Each user could make up to 15 calls, and each call was randomly routed to one of the available systems. A total of 274 dialogues were collected from 26 users after discarding those who did not speak to all three systems. The results are shown in Table 3. None of the differences is statistically significant. Moreover, the MOS responses (Overall, Emphasis, and Intonation) are highly correlated to each other (\( r > 0.8 \)) and moderately correlated to Comprehension (0.40, 0.29, 0.33).

<table>
<thead>
<tr>
<th></th>
<th>Voice</th>
<th>Success</th>
<th>Compr</th>
<th>Overall</th>
<th>Emphasis</th>
<th>Intonation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>91.9%</td>
<td>3.65</td>
<td>42.6</td>
<td>41.3</td>
<td>41.8</td>
<td></td>
</tr>
<tr>
<td>Emphatic</td>
<td>90.3%</td>
<td>3.85</td>
<td>42.2</td>
<td>40.6</td>
<td>40.5</td>
<td></td>
</tr>
<tr>
<td>Dialogue</td>
<td>89.0%</td>
<td>3.78</td>
<td>42.4</td>
<td>41.7</td>
<td>41.7</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Summary of the user trial results. Each row corresponds to a synthetic voice and each column to the question asked.

Analysis of Variance (ANOVA) was performed on the data, in order to discover which factors affected the users’ responses. Table 4 summarizes the one-way ANOVA results for each of the MOS answers against different factors. The results show that Comprehension is the most significant factor in explaining the variance for all the MOS observations. Success, having moderate correlation with Comprehension (0.38), is marginally a significant factor. On the other hand, the Voice factor has no significant effect on any of the Overall, Emphasis, or Intonation MOS factors. Two-way and three-way ANOVA was also performed, however no significant effect was found for any combination involving Voice factor. The results show that the design of the experiment was not effective in assessing the utility of the synthetic voices. The users could not disentangle the primary task of maintaining the dialogue to find a venue from the secondary task of evaluating the quality of the synthetic speech.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Overall</th>
<th>Emphasis</th>
<th>Intonation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice</td>
<td>0.02 (0.971)</td>
<td>0.16 (0.851)</td>
<td>0.22 (0.801)</td>
</tr>
<tr>
<td>Success</td>
<td>5.45 (0.020)</td>
<td>1.07 (0.300)</td>
<td>3.84 (0.051)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>47.4 (&lt;0.001)</td>
<td>22.2 (&lt;0.001)</td>
<td>31.2 (&lt;0.001)</td>
</tr>
</tbody>
</table>

Table 4. One-way ANOVA results of the MOS answers compared to the Voice, Success, and Comprehension factors. The F-value is shown for each pair as well as, the significance level (p-value).
3.3. Listening Test

Given the above, a preference listening test was designed to evaluate the three voice setups in the context of a spoken dialogue system. The listener was presented a dialogue script including both the system prompts and the user responses. The top ASR hypothesis was used as the user response instead of the actual user’s speech transcription so that the listener is not affected by any misrecognitions. Each system turn had a pair of alternative synthetic prompts, and the listener was asked to choose the most appropriate one or indicate no preference. The presentation order of the two alternative prompts was randomized. One could listen to each pair multiple times, though this happens rarely with crowd-sourced evaluators.

The voices were evaluated in pairs. A set of 50 dialogues were randomly selected from the ones collected during the user trial. The system prompts were synthesized with all the three voice setups, using the actual dialogue acts and the emphasis tags that were assigned at runtime. For each dialogue three listening tasks were generated (one per comparison). Each task was evaluated at most 6 times via crowd-sourcing. A total of 339 evaluators completed the listening task, resulting in a total of 6395 judgments.

The results are shown in Table 5 and are organized in three sections. The top section compares the baseline voice versus the emphatic voice, the next section compares the emphatic voice to the dialogue context-sensitive voice, and the last section compares the baseline to the context-sensitive voice. For each comparison, the total preference percentages are shown, as well as the breakdown according to conditions. The first one is whether the prompt contained an emphasized slot (emphasis) or not (plain), while the other breaks down the results according to the dialogue act type (confirm - the system is confirming a slot, confreq - confirming a slot while requesting another, inform - informing one or more slots, and request - requesting information for a slot). The number of judgements per comparison is also shown, as well as the statistical significance level estimated using a sign test.

The comparison between the baseline and the emphatic voice shows significant preference towards the emphatic one. This preference is mainly attributed to the sentences containing emphasized slots, while there is insignificant preference to the baseline voice in case of prompts without emphasis (plain). The preference is also significant for the inform dialogue act. This is expected since more than half of the total number of prompts were of inform type and about half of them contained emphasized slots. The comparison between the emphatic voice and the context-sensitive one shows significant preference towards the latter, when there is no emphasis present, while there is no preference otherwise. Moreover, the latter is more preferable for all the different dialogue act types (significantly for the confirm and request types). The final comparison shows significant preference for the context-sensitive voice compared to the baseline regardless of the emphasis presence. The preference is significant for the inform and confreq dialogue acts.

3.4. Discussion

The results largely agree with the intuition given the training setup. The emphasis factor makes a difference only for emphasized sentences (both emphatic and context-sensitive voices are significantly preferred to the baseline), otherwise there is no effect (no significant difference between baseline and emphatic, while the preference to the context-sensitive over the baseline is attributed to the dialogue context). Note that the emphasis features are correlated with some of the baseline features, e.g. content word or accent features [21], so the baseline voice can produce emphatic speech to some extent, ex-
5. REFERENCES


