Predicting how it sounds: Re-ranking dialogue prompts based on TTS quality for adaptive Spoken Dialogue Systems

Cedric Boidin\textsuperscript{1}, Verena Rieser\textsuperscript{2}, Lonneke van der Plas\textsuperscript{3}, Oliver Lemon\textsuperscript{2}, Jonathan Chevelu\textsuperscript{1}

\textsuperscript{1}Orange Labs, Lannion, France
\textsuperscript{2}School of Informatics, University of Edinburgh, UK
\textsuperscript{3}Department of Linguistics, University of Geneva, CH

\texttt{cedric.boidin@orange-ftgroup.com, vrieser@inf.ed.ac.uk}

Abstract

This paper presents a method for adaptively re-ranking paraphrases in a Spoken Dialogue System (SDS) according to their predicted Text To Speech (TTS) quality. We collect data under 4 different conditions and extract a rich feature set of 55 TTS runtime features. We build predictive models of user ratings using linear regression with latent variables. We then show that these models transfer to a more specific target domain on a separate test set. All our models significantly outperform a random baseline. Our best performing model reaches the same performance as reported by previous work, but it requires 75\% less annotated training data. The TTS re-ranking model is part of an end-to-end statistical architecture for Spoken Dialogue Systems developed by the ECFP7 CLASS\textsuperscript{C} project.

Index Terms: speech synthesis, spoken dialogue systems, TTS quality prediction, re-ranking.

1. Introduction

Ultimately, one would like to know “how good it will sound” before generating a prompt in a Spoken Dialogue System (SDS). Evidence from a corpus collected by \cite{1} shows the importance of considering Text To Speech (TTS) quality for SDS: 5.2\% of the user utterances indicate a problem with the TTS quality. For example, the user asks the system to repeat because s/he was not able to understand what the system said.

In this paper we present a re-ranker model to select paraphrases that are predicted to sound most natural when being synthesised with unit selection TTS, following previous work by \cite{2, 3}. However, our approach requires 75\% less annotated data, while reaching the same prediction performance. We first gather training data on user ratings of synthesised prompts (Section 2). We then use linear regression with latent variables to predict the perceived user rating (Section 3). Finally, we evaluate the model on a separate test corpus of paraphrases of possible system prompts (Section 4). This re-ranker model is used in the general architecture of the CLASS\textsuperscript{C} project (see figure 1).

In this project we propose a end-to-end statistical treatment of uncertainty and context adaptive strategies for Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Dialogue Management (DM), Natural Language Generation (NLG), and TTS. In this framework, the re-ranker chooses between alternative inputs to the TTS module according to text-to-speech module capabilities. These alternative inputs are constructed by the statistical NLG module of the CLASS\textsuperscript{C} architecture \cite{4}. The target domain is Interactive Voice Response (IVR) applications, especially troubleshooting and customer service domains (see for example \cite{5}). Ultimately, this work should lead to more robust SDS with higher user ratings, because the system will be able to select its best possible output based on an accurate predictive model of users’ TTS ratings.

2. Training data collection

2.1. Synthesised utterances

We first collect training data. 144 different utterances were synthesised with a female French voice of the state-of-the-art unit selection Orange Labs speech synthesiser. \footnote{See demonstrator at \url{http://tts.elibel.tm.fr}}. The utterances were taken from two application domains: IVR applications (e.g. \textit{“J’annule votre demande.”}– I cancel your order), and movie subtitles (e.g. \textit{“Tu me donnes quel âge ?”}– How old do you think I am?). We choose the training data to represent a wider spectrum than just the CLASS\textsuperscript{C} target domain, in order to build a more general model, which has the potential to transfer to different settings. Furthermore, each utterance was synthesised with two different versions of the TTS voice. One version uses the full inventory (approximately 3 hours of speech), and the other one uses a reduced inventory (30\% of the full inventory, approximately 1 hour of speech, selected randomly). Thus, the training set contains utterances from 4 different conditions, contributing equal portions to the training set (n=288); IVR/full are the IVR utterances synthesised with the full acoustic inventory, IVR/red are the IVR utterances that are synthesised with the reduced inventory, ST/full are the subtitle utterances synthesised with the full inventory and ST/red are the subtitle utterances synthesised with the reduced inventory.
2.2. Subjective evaluation

12 French native listeners were each asked to rate the speech quality of 48 synthesised utterances on a MOS (Mean Opinion Score) 1-5 scale, 1 for bad to 5 for excellent, resulting in 288 utterances for training. For comparison, [3] used a data set of 1236 utterances for training (ca. 75% more).

The MOS for the whole training data (all), IVR/full, IVR/red, ST/full and ST/red are displayed in figure 2 with 95% confidence intervals. A Wilcoxon signed rank test shows that IVR is rated significantly (p < .001) higher than ST, and the full version is rated significantly higher (p < .001) than the red version. Furthermore, a pair-wise comparison off all the 4 conditions (with Bonferroni correction) shows that IVR/full significantly (p = .006) outperforms all the other conditions. These results illustrate the fact that the chosen voice is mainly tailored to the IVR domain, and that the inventory size of the voice plays a significant role in the perceived quality of synthesised speech.

2.3. Feature extraction

A set of 55 runtime features was automatically extracted for each synthesised utterance. Most of the features reflect the contents of the acoustic inventory used to build the voices. Language models (LM) were constructed using the SRILM toolkit. Several features were extracted for each LM: coverage rate, number of Out-Of-Vocabulary words, perplexity measure and log-probability measure. The features are of several types:

- The unit selection cost of the speech synthesiser.
- 1-4 gram phoneme level LMs.
- 1-2 gram “Sandwich unit” LMs: “Sandwich units” are a new phonetic units, that reflect properties of the unit selection algorithm, and thus are presumed to be related to speech quality. “Sandwich units” are defined as phonetic strings composed of one or more consecutive vow-

2.4. Correlations

We first analyse the correlation between each individual feature and the MOS ratings for the whole data set. Table 1 contains the top 10 features ranked by the absolute value of their correlation with the MOS ratings. The maximum correlation is +0.57, which is obtained for the coverage rate of the poor sandwich 2-grams. 6 out the 10 top features concern the sandwich units, showing the validity of this new unit. A high number of features show a significant correlation with the MOS ratings, although none of them reaches 0.6. We also observe a lot of redundancy between the different features. We deliberately choose a large number of features, in order to detect the most relevant ones for predictive modelling, as described in the next section. Note that none of the previous work has used such a rich feature set: We substantially extend the feature set used by [3] by using 10 times more features.

![Figure 2: Mean Opinion Scores for the training data: IVR/full (3.99), IVR/red (3.52), ST/full (3.22) and ST/red (2.89), all (3.40).](http://urd.let.rug.nl/tiedeman/OPUS/OpenSubtitles.php)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>coverage rate of poor sandwich 2-grams</td>
<td>+0.57</td>
</tr>
<tr>
<td>coverage rate of rich sandwich 2-grams</td>
<td>+0.53</td>
</tr>
<tr>
<td>unit selection cost</td>
<td>-0.49</td>
</tr>
<tr>
<td>coverage rate of rich sandwich units</td>
<td>+0.47</td>
</tr>
<tr>
<td>coverage rate of poor sandwich units</td>
<td>+0.44</td>
</tr>
<tr>
<td>number of Out-Of-Vocabulary rich sandwches</td>
<td>-0.44</td>
</tr>
<tr>
<td>number of Out-Of-Vocabulary poor sandwiches</td>
<td>-0.41</td>
</tr>
<tr>
<td>number of Out-Of-Vocabulary words</td>
<td>-0.39</td>
</tr>
<tr>
<td>perplexity of phoneme 3-grams</td>
<td>-0.39</td>
</tr>
<tr>
<td>coverage rate of words</td>
<td>+0.37</td>
</tr>
</tbody>
</table>

Table 1: The top 10 features for their correlation with MOS

3. Predictive modelling

We use stepwise linear regression to build predictive models, following the general framework of [6]. Stepwise regression automatically selects the most relevant features, i.e. the ones which are most predictive of the average user rating for each utterance. We experiment with 3 different models (see Table 2): simple regression, regression using latent variables, and a reduced version of latent variables. We use data fit (R²adj) for assessing the quality of the model with Leave-One-Out cross-validation.

The first model is a simple linear regression model, using all the features as input variables. However, an analysis of the

---

1 A one-way ANOVA with pair-wise PostHoc analysis (Bonferroni correction) shows that some users rate significantly different than other users (p = .000). However, when comparing the 2 user ratings on the same utterance, the ratings are highly correlated (Pearson’s r=.504, p=.000), which indicates that there is a general agreement on the quality of one sentence.

2 http://www.speech.sri.com/projects/srilm/

[^3]: http://urd.let.rug.nl/tiedeman/OPUS/OpenSubtitles.php
weight | predictor
--- | ---
+ .294 | 2-gram poor sandwich unit coverage rate
- .175 | number of Out Of Vocabulary words
- .215 | unit selection cost
+ .125 | perplexity for language model 1-gram on phonemes
- .110 | perplexity for language model 1-gram poor sandwich units

Simple Regression model: \( R^2_{\text{adjusted}} = .397; \)

FA model: \( R^2_{\text{adjusted}} = .363; \)

FA3: \(.774^*\text{perplexity for 3-gram LM on words} + .735^*\text{unit selection cost}+\ldots\)
FA4: \(.870^*\text{perplexity for 2-gram LM on phonemes} + .859^*\text{perplexity for 4-gram LM on phonemes}+\ldots\)
FA2: \(.971^*\text{perplexity for 3-gram LM on words} + .972^*\text{perplexity for 4-gram LM on words}+\ldots\)
FA6: \(.944^*\text{perplexity for 1-gram LM on phonemes}+\ldots\)
FA7: \(.510^*\text{words coverage rate} + .463^*\text{perplexity for 1-gram poor Sandwich Unit}+\ldots\)
FA5: \(.909^*\text{perplexity for 2-gram poor Sandwich Unit} + .898^*\text{perplexity 2-gram rich Sandwich Unit}+\ldots\)

Reduced FA model: \( R^2_{\text{adjusted}} = .345; \)

| weight | predictor |
--- | --- |
- .38 | unit selection cost |
- .05 | perplexity for 4-gram LM on phonemes |
- .03 | perplexity for 3-gram LM on words |
+ .19 | perplexity for 1-gram LM on phonemes |
+ .25 | words coverage rate |
- .07 | perplexity for 2-gram poor Sandwich Unit |

Table 2: Linear regression models to predict average user rating of synthesised prompts.

4. Evaluation

4.1. Test data collection

We first investigate whether the model built from the more general training corpus transfers to the IVR/trouble-shooting domain, which we target in the CLASSIC project. Furthermore, we are interested how well the different regression models perform for re-ranking paraphrases (so far we only tested prediction accuracy of single utterances). Therefore, we collected a separate test set in the IVR domain with user ratings of 20 sets of 3 paraphrases each (n=60):

- A first utterance was directly taken from a prototype of a commercial application [5], e.g. “Pourriez-vous me donner la référence de votre modem?”.
- A second utterance was produced by replacing words in the first utterance with their synonyms, e.g. “Pourriez-vous me préciser la référence de votre modem?”; Synonyms were acquired automatically from the EuroParl corpus using the alignment-based distributional method decribed in [7] and selected for substitution by hand.
- A third utterance was produced with an automatic paraphrase generator using bilingual parallel corpora, following a similar method to [8], e.g. “Pourriez-vous me donner la référence de votre modem?”.

The generated utterances were manually edited to prevent worse paraphrases from influencing the user ratings. The utterances are synthesised with the voice based on the full acoustic inventory. This version will be used in the project, as the MOS scores were significantly higher (see section 2.4). 5 French native speakers were asked to rate the 20 sets, on a 1-5 scale MOS scale, where each set was presented as a list of 3 paraphrases.

4.2. Evaluation

The three regression models are evaluated on the test data and compared against a random baseline, following [3]. The “1-best” re-ranking task consists of predicting the utterance which
is ranked first by humans. The “pairs” re-ranking task consists of predicting the ordering for each possible pair of paraphrases. The “sets” re-ranking task consists of predicting the ordering for each set of 3 paraphrases. The “1-best MOS” performance is the average MOS of the utterances that are ranked first, which can reach 3.99 at the maximum. The 1-best MOS gain is the relative improvement in MOS over the random baseline. Table 3 shows the performance of the 3 regression models on the test data. The average MOS rating for the test data is 3.30(±.79). The performances for the 1-best, pairs and sets re-ranking models are compared to the random baseline using a one-tailed binomial test. The MOS of the 1-best predicted utterance is compared to the random baseline using a one-tailed paired t-test. For all these measures, the three regression models perform significantly better than the random baseline at a significance level of 95% (see detailed p-values in table 3). The simple regression model performs slightly worse than the Factor Analysis model and the reduced FA model. However, no significant difference between the three regression models can be found. The Factor Analysis regression model performs best (although not significantly), and reaches performance levels that are equivalent to the best model of [3]: 78% of pair ordering and 66% of possible MOS gain.

4.3. Error Analysis

The MOS RMS errors of the different models are relatively high (about 0.8). A closer analysis of the results shows that the predicted MOS scores are consistently higher (on average 0.5 point) than the actual user ratings. A possible explanation for this difference is that different user populations were used for data collection: users in the training data collection gave on average higher scores for IVR utterances and full inventory (MOS 3.99) than users in the evaluation data collection (MOS 3.30), which reflects the difference for the predicted MOS scores. Users in the training data collection might have rated the IVR utterances better, as they were also exposed to synthesised subtitles, where the TTS quality was worse (see Section 2.2).

5. Conclusion and Discussion

This paper presents a method for re-ranking paraphrases in a Spoken Dialogue System (SDS) according to their predicted Text To Speech (TTS) quality. We collected data (n=288) under 4 different conditions (2 domains, 2 versions of a TTS voice) and extracted a rich feature set of 55 TTS runtime features. We built predictive models of user ratings using linear regression with latent variables. We then showed that these models transfer to a more specific target domain (n=60). All our models significantly outperform a random baseline. Our best performing model reaches the same performance as reported by [3]. In contrast to [3] we use a much smaller data set for training (it requires 75% less annotated data), while using a larger set of features (10 times more).

The TTS re-ranking model is part of an end-to-end statistical architecture developed by the CLASSIC project (www.classic-project.org). The ultimate goal of this work is to contribute to more robust and adaptive Spoken Dialogue Systems with higher user ratings. This system is able to reliably select a best possible output based on an accurate predictive model of users’ TTS ratings.

In future work we will test the re-ranker performance when integrated in a full working system. This will allow us to investigate differences between over-hearer ratings of played-back utterances vs. interactive users engaged in a real dialogue.

6. Acknowledgments

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216594 (CLASSIC project: www.classic-project.org).

7. References


Table 3: Re-ranker accuracy (correct/total), MOS root mean square (RMS) error, 1-best MOS and 1-best MOS gain of the 3 regression models and their p-values compared to the random baseline.

<table>
<thead>
<tr>
<th>Number of comparisons</th>
<th>Random baseline</th>
<th>Simple regression</th>
<th>FA regression</th>
<th>Reduced FA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-best re-ranking</td>
<td>20</td>
<td>33%</td>
<td>60% (p = 0.012)</td>
<td>70% (p = 7.8e-4)</td>
</tr>
<tr>
<td>Pairs re-ranking</td>
<td>60</td>
<td>50%</td>
<td>70% (p = 0.001)</td>
<td>80% (p = 1.6e-6)</td>
</tr>
<tr>
<td>Sets re-ranking</td>
<td>20</td>
<td>17%</td>
<td>40% (p = 0.013)</td>
<td>55% (p = 1.3e-4)</td>
</tr>
<tr>
<td>MOS RMS error</td>
<td>60</td>
<td>0.87</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td>1-best MOS</td>
<td>20</td>
<td>3.30</td>
<td>3.60 (p = 0.014)</td>
<td>3.77 (p = 5.2e-4)</td>
</tr>
<tr>
<td>1-best MOS gain</td>
<td>30</td>
<td>0%</td>
<td>44%</td>
<td>68%</td>
</tr>
</tbody>
</table>