A Phonetic Adaptation Module for Spoken Dialogue Systems

Eran Raveh
Multimodal Computing and Interaction
Computational Linguistics & Phonetics
Saarland University
raveh@coli.uni-saarland.de

Ingmar Steiner
DFKI GmbH, Saarbrücken
Computational Linguistics & Phonetics
Saarland University
steiner@coli.uni-saarland.de

Abstract

This paper presents a novel component for spoken dialogue systems, which adds the functionality of adapting the system's speech output based on the user's input. The adaptation in done on the phonetic level for adopting the user's speech characteristics without changing the system's own voice. An architecture for a spoken dialogue system is introduced, in which this module creates a direct link between the speech recognition and the speech synthesis modules.

1 Introduction

In a typical workflow of a spoken dialogue system (SDS), the automatic speech recognition (ASR) and the text to speech (TTS) modules work separately, meaning that the speech input and output are completely unrelated and function merely as speech-to-text and text-to-speech transformers. This means that a system's output will be pronounced in the same manner, regardless of how the user speaks to it. The module implementation introduced in this paper aims to create a more direct connection between the ASR and the TTS modules. Such a connection enables the direct influence of the user's input on the system's output on the phonetic level.

Such adaptation (or convergence) capabilities make it possible for the system to personalize its output to the user's style of speech. Seeing that convergence between interlocutors occurs in human-human interaction (Pardo et al., 2010), triggering it in human-computer interaction may lead to a more natural – and therefore more fluent – interaction. This feature can be beneficial, among others, for social chatbots, for their main purpose is to create a natural and personalized interaction with the user. More specific applications could utilize it for more goal-driven tasks, like pronunciation tutoring or capturing dialectal differences.

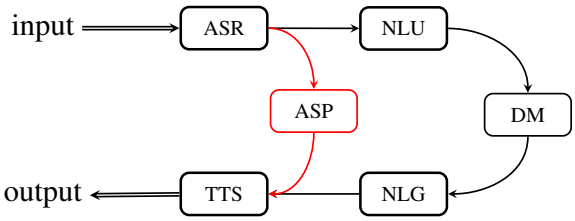

Figure 1: Architecture of an SDS with an additional component and connections (in red) between the ASR and TTS components, which performs additional speech processing for phonetic adaptation.

2 System

We present here an end-to-end dialogue SDS with an additional module that supports phonetic adaptation (see Figure 1).

2.1 Architecture

In this work, OpenDial framework (Lison and Kennington, 2016) was used for creating a modular spoken dialogue system architecture. Some of its built-in components were used, including the natural language understanding (NLU), dialogue manager (DM), and natural language generation (NLG) modules. A new ASR module was implemented, which includes some additional functionality for detecting the target segments and extracting relevant metadata to pass to the ASP module (see below). A new TTS module was also implemented, using Praat[1] as the signal processing back-end. This module is needed for the transformation of the phonetic data output of the ASP

[1] http://www.fon.hum.uva.nl/praat/

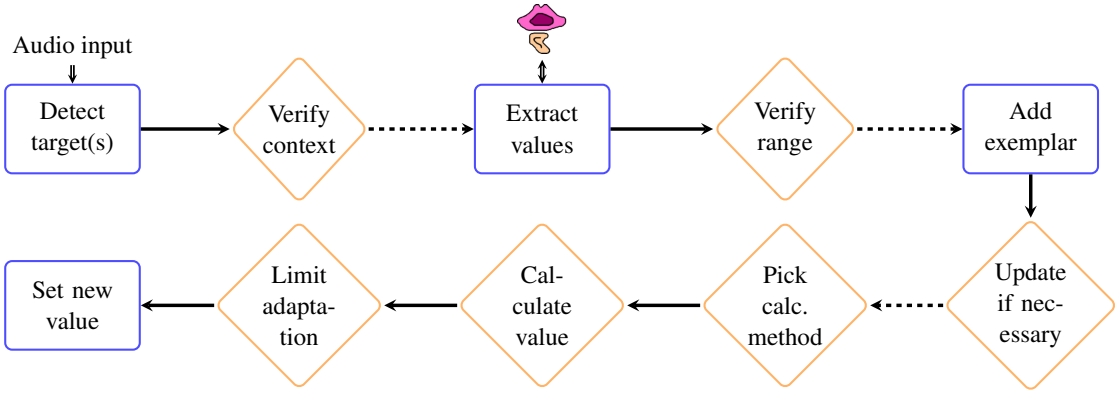

Figure 2: Overview of the adaptation pipeline integrated into the ASP module, with Praat as the signal processing back-end. Mandatory, fixed steps are marked by blue rectangles and parameterized steps by orange diamonds. Dashed arrows mark conditional transitions that terminate the process if they are not fulfilled. All the steps are explained in detail in Raveh et al. (2017).

module into articulation properties. The main addition to the typical SDS model is the additional speech processing (ASP) module. This module extracts phonetic features from the speech signal and ASR output, and provides their adapted values to the TTS module, where the adaptation is realized in the synthetic speech. These values are the output of the pipeline presented in Raveh and Steiner (2017). The flow of this pipeline is summarized in Figure 2. This module takes input which combines some customized functionalities of the ASR module and the feature tracking and adaption pipeline.

2.2 Models

A subset of the system's modules contribute to its response to the user. To sum up, the DM module determines *why* the output utterance it generated, the NLG module *what* will be uttered, and finally the ASP module defined *how* it will be uttered. We created a new XML-based OpenDial domain with simple NLU and NLG models using manually crafted rules for handling user intent and system response. The ASR component uses standard Voxforge[2] acoustic models for German dictionary and language model designed especially for this system. The segment-level adaptation is realized through the phonetic response model introduced in Raveh et al. (2017). This model adapts to given input speech on the segmental level. The goal of the model is to adapt to the user's speech *characteristics*, while avoiding changes in the voice itself. The adaptation behavior can be modified by various parameters, e.g., allowed value range, update frequency, convergence rate, convergence limit, and more. These parameters are a computational representation of behavior observed in human-human interaction while listening to synthetic stimuli.

[2] `http://www.voxforge.org/de/downloads`

3 Summary

A novel module for adding phonetic adaptation capabilities to SDSs based on a computational convergence model was presented. This module was integrated into an end-to-end SDS, making it possible for the phonetic characteristics of the system's output to be adapted to those of the user. Future work includes using this architecture for a task-specific system to evaluate such adaptation and its effect on the user's behavior.

References

Pierre Lison and Casey Kennington. 2016. OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *ACL: System Demonstrations*, pages 67–72.

Jennifer S. Pardo, Isabel Cajori Jay, and Robert M. Krauss. 2010. Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, 72(8):2254–2264.

Eran Raveh and Ingmar Steiner. 2017. Real-time pipeline for segmental feature tracking and adaptation with Praat. In *Phonetik und Phonologie*. accepted.

Eran Raveh, Ingmar Steiner, and Bernd Möbius. 2017. A computational model for phonetically responsive spoken dialogue systems. In *Interspeech*. in press.