

Speech Structure Acquisition for Interactive Systems

Holger Brandl^{1,3}, Miguel Vaz^{2,3},
Frank Joublin³ & Christian Goerick³

¹ Research Institute for Cognition and Robotics, Bielefeld University, Germany

² Department for Industrial Electronics, Universidade do Minho, Portugal

³ Honda Research Institute Europe GmbH, Offenbach am Main, Germany

`hbrandl@techfak.uni-bielefeld.de`

INTRODUCTION

Robots in social interaction need to be able to communicate with humans in a human-like manner by understanding and using their language. One important first element of speech and language understanding is the ability to parse spoken utterances into words. But this ability is not innate and needs to be developed by infants within the first years of their life. So far almost all computational speech processing systems neglected this bootstrapping process. Here we propose a model for early infant word learning embedded into a layered architecture comprising phone, phonotactics and syllable learning. Our model uses raw acoustic speech as input and aims to learn the structure of speech unsupervised on different levels of granularity.

In addition to previous work (Brandl *et al.*, 2008) we describe how our model can be used to learn a basic syntax model in interaction. To close the loop between speech structure acquisition and production we further outline how our system can be combined with a speech imitation system that was already described in Heckmann *et al.* (2008) and Vaz *et al.* (2008).

We present first experiments which evaluate our integrated speech acquisition and production model on speech corpora that have some of the properties of infant-directed speech.

SPEECH STRUCTURE ACQUISITION

In order to build a system that is able to learn words based on developmental speech acquisition principles like phonotactically constrained syllable parsing, subtraction-learning, metric segmentation or transitional probabilistic modeling (cf. Aslin *et al.*, 1998; Jusczyk & Peter, 1999), we proposed a three-layered framework for speech acquisition in Brandl *et al.* (2008). Its first layer learns a phone-representation including a phonotactic model. The second layer learns a syllable representation based on the syllabic constraints implied by these learned phonotactics and input speech obeying some properties of infant-directed speech. Finally, our framework acquires a word lexicon based on the above mentioned acquisition principles which we think to play a critical role in early-infant language development. Our system is implemented as a cascade of HMM-based speech unit spotting instances that rely on incomplete speech unit representations on phone, syllable and word level.

SYNTAX MODELING AND UTTERANCE GENERATION

Every structure acquisition layer incrementally bootstraps a bi-gram model. The basic motivation for such a model was to improve recognition rates. The word bigram model can be additionally used in a generative manner. This is a first step to making our system capable of generating syntactically correct utterances, as to what word order concerns. More high-level issues like verbal actions are not touched at all here.

In Mikhailova *et al.* (2008) we described an embodied system running on Honda's Asimo robot where it learns to associate different acoustic labels to various object properties like color, planarity or position. This system is able to learn several synonyms for one semantic concept. Whereas the original system was only able to recognize acquired labels our motivation since then has been to extend such it with the ability to describe a presented object. *E.g.*, given a red apple presented on the right side our system should be able to provide an acoustic scene description like "right red apple".

Given a set of active semantic categories we propose a graph theoretic search algorithm that is based on the acquired bi-gram models, a non-empty set of synonyms for each semantic concept and optional cues like current interaction language to determine a syntactically correct word symbol sequence.

VIRTUAL SPEECH IMITATION FOR PRODUCTION

Whereas a direct playback of the corresponding speech snippets (collected during training of the word models) could be easily implemented, we do not consider this to be usable for intuitive human robot interaction. Neither is there a difference between the voice of the robot and its tutor, nor is it possible to change articulation parameters like timbre or pitch as required by the state of interaction.

We implement a more sophisticated scheme by coupling our speech acquisition framework with the speech mimicking system described in Heckmann *et al.* (2008). This system reproduces an arbitrary word or utterance by using a resynthesis scheme which allows to alter pitch, formant and spectral structure of the input speech. Although our coupling is done using a simple virtual playback approach with subsequent resynthesis, we consider this nevertheless to be an important step towards embodied online learning of speech and language abilities.

REFERENCES

- Aslin, Richard N., Saffran, Jenny R., & Newport, Elissa L., 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4):321-324.
- Brandl, Holger, Joublin, Frank, Wrede, Britta, & Goerick, Christian, 2008. A self-referential childlike model to acquire phones, syllables and words from acoustic speech. In Accepted for *7th International Conference on Development and Learning*.
- Heckmann, Martin, Glaeser, Claudius, Vaz, Miguel, Rodemann, Tobias, Joublin, Frank, & Goerick, Christian, 2008. Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.
- Jusczyk, Peter W., 1999. How infants begin to extract words from speech. *Trends in Cognitive Sciences*, 3(9):323-328.
- Mikhailova, Inna, Heracles, Martin, Bolder, Bram, Janssen, Herbert, Brandl, Holger, Schmüdderich, Jens, & Goerick, Christian, 2008. Coupling of mental concepts to a reactive system: incremental approach in system design. In Submitted to the *Eighth International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*.
- Vaz, Miguel, Brandl, Holger, Joublin, Frank, & Goerick, Christian, 2008. A computational model for early infant speech acquisition and imitation. In *Workshop about Speech and Face to Face communication*.