

# Linking Perception and Production: System Learns a Correspondence Between its Own Voice and the Tutor's

Miguel Vaz<sup>1,3</sup>, Holger Brandl<sup>2,3</sup>, Frank Joublin<sup>3</sup> & Christian Goerick<sup>3</sup>

<sup>1</sup> Universidade do Minho, Portugal

<sup>2</sup> Universität Bielefeld, Germany

<sup>3</sup> Honda Research Institute Europe GmbH, Germany

mvaz@dei.uminho.pt; hbrandl@techfak.uni-bielefeld.de;  
Frank.Joublin@honda-ri.de; christian.goerick@honda-ri.de

## ABSTRACT

We hereby present our first steps towards linking an embodied speech acquisition system and a speech production module, in order to provide the system with the ability to produce acquired speech representations. Due to the type of interaction planned with the system, we endowed it with a child-like voice, concretized with the use of a vocoder-like technique for speech synthesis. The task in hand consists of finding and using a correspondence between configurations in the tutor's acoustic parameter space, which might be untangible for the system's voice, and phonologically equivalents in the robot's.

## INTRODUCTION

The ability to use natural spoken language to interact with a robotic system like Honda's ASIMO is highly desirable because it greatly increases the naturalness and efficiency of the communication. Ideally, in order to face the demand for flexibility of such a task, the acquisition process would make little or no assumptions about the used language and adapt itself to the characteristics relevant in the environment, much like children do in the first years of their lives. The speech representation should be learned during the interaction with a tutor instead of being predefined.

In a direct interaction context, it is also desirable that the system creates no false expectations about its real capabilities. For this reason, we decided that the system should speak with a child's voice. A child's voice is more appropriate for the type of interaction that takes place in a learning system like (Schmüdderich *et al.*, 2008), where almost no knowledge of the world is assumed. Furthermore, it provides a framework in which it can be tested how children solve the correspondence problem in imitation learning: transfer the relevant perceptual auditory features of the utterances of their parents into acoustic goals which are attainable by their own different vocal tract.

From the perceptual point of view, our first steps towards a system that fulfills the aforementioned requirements have already been reported in Brandl *et al.* (2008). There, it has been shown how phones, syllables and words can be learned in an unsupervised fashion using a child-directed speech. These speech recognition capabilities have already been integrated in an autonomous learning interaction framework working on the humanoid robot ASIMO, where it could acquire speech labels for objects and attributes of objects, like size, position and orientation (Schmüdderich *et al.*, 2008). Thanks to these previous works, ASIMO is able to recognize previously learned speech-labels.

It is, however, not yet possible for it to produce an audible description of a given scene or object using the same previously learned representations. We hereby attempt to fill this gap by combining the already mentioned acquisition system with a speech production model.

## AUDITORY-MOTOR MAPPING

Unlike in conventional ASR systems, the acquired speech representations are not based on symbolic phoneme sequences. As a consequence, we cannot base the perception/production link on such symbolic representations, like TTS systems do. Instead, we learn and use a mapping between

acoustic feature spaces of the tutor's and the system's voice. Due to their known perceptual relevance, we use formant frequencies as perceptual features. With this mapping, it is possible for the system to represent an input utterance as a trajectory in its formant space, and use the projected trajectory to synthesize the same utterance with its own voice. At the moment, the mapping takes the technical form of a K-Nearest Neighbours, trained with the imitation response of the tutor to a set of utterances from the robot. We share the belief with other researchers (Miura *et al.*, 2007), that small children are imitated by their caregivers and that this imitation provides them with necessary information for grounding their voice into that of their tutor's.

## **SYNTHESIS TECHNIQUE**

As opposed to most of the work done in similar research (Guenther & Perkell, 2004; Miura *et al.*, 2007), where articulatory production models are used, we use an acoustic production model with a vocoder-like architecture. We use spectral features to synthesize speech, extracted using a gammatone filterbank, which has an optimal tradeoff between spectral and temporal resolution. As a consequence, we can synthesize high (women's and children's) and low (men's) pitch voices (when compared to the synthesis with standard features), with similar naturalness and intelligibility.

The child's voice is grounded in a set of spectral vectors that we use as *motor primitives*. These are, in a first step, taken per hand from utterances from the model child, although it is foreseeable to use a method like that described in Miura *et al.* (2007) to guide their acquisition. Each of the spectral vectors was chosen so that it corresponds to a vowel. In order to combine the motor primitives, and to generate smooth transitions between them, we introduce a spectral morphing algorithm that works with by the trajectories of the formants. Using these projected trajectories, we are able to very accurately imitate completely voiced utterances (vowel and diphthong with the highest confidence) from the tutor with the system's child voice. Unvoiced segments are, however, representable in a limited way. Nevertheless, the vocabulary that is used in a session with the interactive system is also limited: to test the appropriateness of the aforementioned system coupling, we evaluate the intelligibility of the production with a set of words acquired by the robot in a learning session. While remaining a reactive solution, we are able to extend our existing parrot-like real-time speech imitation system (Heckmann *et al.*, 2008).

## **BIBLIOGRAPHY**

- Brandl, H., Joublin, F., Wrede, B. & Goerick, C., 2008. A self-referential childlike model to acquire phones, syllables and words from acoustic speech, 7th International Conference on Development and Learning.
- Guenther, F.H. & Perkell, J.S., 2004. A neural model of speech production and its application to studies of the role of auditory feedback in speech. In: B. Maassen, R. Kent, H. Peters, P. Van Lieshout, and W. Hulstijn (eds.), *Speech Motor Control in Normal and Disordered Speech* (pp. 29-49). Oxford: Oxford University Press.
- Heckmann, M., Glaeser, C., Vaz, M., Rodemann, T., Joublin, F. & Goerick, C., 2008. Listen to the Parrot: Demonstrating the Quality of Online Pitch and Formant Extraction Via Feature-Based Resynthesis, Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2008.
- Miura, K., Yoshikawa, Y. & Asada, M., 2007. Unconscious Anchoring in Maternal Imitation that Helps Finding the Correspondence of Caregiver's Vowel Categories, *Advanced Robotics*, 21(13), 1583-1600(18).
- Schmüdderich, J., Brandl, H., Bolder, B., Heracles, M., Janssen, H., Mikhailova, I. & Goerick, C., 2008. Organizing Multimodal Perception for Autonomous Learning and Interactive Systems, submitted to IEEE-RAS International Conference on Humanoid Robots 2008.
- Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W. & Tokuda, K., 2007. *The HMM-based speech synthesis system version 2.0*, Proc. of ISCA SSW6, Bonn, Germany, Aug. 2007.