Language Acquisition Embedded into Tutor-Robot Interaction

Martin Heckmann, Holger Brandl, Xavier Domont, Miguel Vaz, Jens Schmüdderich, Bram Bolder, Frank Joublin, Christian Goerick

2009

Language Acquisition Embedded into Tutor-Robot Interaction

Martin Heckmann[1], *Holger Brandl*[1,2], *Xavier Domont*[1,3], *Miguel Vaz*[1,4],
Jens Schmuedderich[1], *Bram Bolder*[1], *Frank Joublin*[1], *Christian Goerick*[1]

[1]Honda Research Institute GmbH, Offenbach/Main, Germany
[2]Research Institute for Cognition and Robotics, University of Bielefeld
[3]Technische Universität Darmstadt, Control Theory and Robotics Lab
[4]University of Minho, Guimaraes, Portugal, Department of Industrial Electronics
firstname.lastname@honda-ri.de, hbrandl@cor-lab.uni-bielefeld.de,
xavier.domont@rtr.tu-darmstadt.de mvaz@dei.uminho.pt

Abstract

Children acquire language to a large extend in the interaction with their caregivers. Inspired by this observation we develop computational models and artifacts for the acquisition of language in an interactive scenario. The artifact bootstraps its representations with little a priori knowledge and can be taught by a human tutor. In this framework we investigate different aspects of the speech acquisition process. This encompasses the learning of speech features, word and sub-word units as well as the production of acquired speech units. As speech features we explore a set of hierarchical spectro-temporal features which are learned in an unsupervised fashion based on the observed speech data. Phone-like speech units emerge from an unsupervised clustering process. These phone-units can then be used to bootstrap word learning in an interactive scenario where a tutor shows a visual property and at the same time utters a corresponding speech label. Thereby an auditory attention system and predefined key phrases trigger the learning behavior. Finally the learned units can also be reproduced.

Index Terms: speech acquisition, speech features, word learning, speech synthesis, attention system

1. Introduction

Common models of spoken language processing decode the utterance based on predefined features, vocabularies, grammars, and knowledge bases. This does not reflect the way children learn language in the interaction with their environment. In their struggle to structure their environment children have to rely on the cues provided by their caregivers and those intrinsic in the statistics of the environment.

To yield a system capable of bootstrapping its representations with little initial knowledge and featuring open ended development we take in our work inspirations from recent findings in developmental psychology an neurobiology. A key feature of our approach is the integration of unsupervised, data-driven learning methods and interactive learning from a tutor.

In the following we will exemplify our methodology with results we obtained on different sub-tasks which are required in a system capable of acquiring language similar to the way children do. First we will briefly describe a set of speech features we developed and which are learned based on the input statistics without a supervision signal. Next we will highlight how we learn sub-word speech units to bootstrap a word learning process. After this we introduce an audio-visual attention mechanism which enables our system to focus during learning on relevant aspects of the environment and neglect others. This is followed by an overview on our interactive system which allows a tutor to teach a humanoid robot visual and auditory clusters. Both visual and auditory clusters start with little a priori knowledge and are learned in interaction. As a robotics platform we use Honda's humanoid robot ASIMO. Finally we will indicate how the system can acquire verbal abilities by learning how to imitate the previously learned speech labels.

2. Speech Features

In most approaches the speech features are predefined and mainly based on Mel Frequency Cepstral Coefficients (MFCCs) [1] or RelAtive SpecTral Perceptual Linear Predictive (RASTA-PLP) features [2]. We currently investigate how features can be learned data-driven. Based on inspirations from neurobiology, i.e. the receptive fields in the mammalian primary auditory cortex, and visual object recognition models we developed an acoustic feature extraction framework. The framework is organized in two hierarchical layers and on each layer spectro-temporal receptive fields are learned in an unsupervised way. When using this framework we see significant improvements for the recognition in noise [3] and could also show that an adaptation of the features to different environments is beneficial [4]. For mammals such task specific plasticity of receptive fields in the auditory cortex seems to play an important role [5]. Given that the learning steps involved in our feature extraction framework are unsupervised such an online adaptation seems also to be possible.

3. Sub-Word Units

The previously described features can serve as a basis to learn speech units on a sub-word level. These speech units can then be combined to larger units as syllables and words.

For learning the sub-word units we follow an approach similar to [6]: In the first step single state Hidden Markov Models (HMMs) are learned from a few minutes of untranscribed speech in an unsupervised clustering process based on the k-means algorithm. A transition matrix of these single state HMMs can be used to estimate the most frequent transitions. The combination of these most frequent transitions yields initial 3 state phone models which are then further refined via Baum Welch training [7].

4. Attention Model

The speech acquisition process we described so far was based on completely unsupervised learning strategies. We did first experiments to extend these approaches to the learning of syllables and phonotactic rules [7]. Alternatively we also investigate the learning of speech and visual clusters in an interactive scenario where a tutor teaches a robot [8, 9].

In an interactive scenario the robot perceives a multitude of stimuli, auditory and visual, at the same time. We therefore investigate attention mechanisms to enable the robot to selectively concentrate on one aspect of the environment while ignoring other things [10]. Models of attention, auditory or visual, typically comprise a stimulus driven bottom-up saliency stage and a top-down modulation to enhance or suppress certain types of stimuli [11, 12]. We integrated these aspects into a system which combines a visual and auditory attention system.

4.1. Visual Attention

Our visual attention system is mainly bottom-up driven and based on the concept of proto-objects. Proto-objects are regions in the visual field that are formed by a common grouping feature, can be tracked over multiple images, and are stabilized both in space and time (see [13] for more details). The visual scene description consists of a (possibly empty) set of possibly interesting entities that are close to the robot, move, are large planes, have a certain color, or any possible combination of these. From the set of proto-objects one is selected for interaction, i.e. ASIMO can point, walk, and gaze towards them.

Mainly objects in the peri-personal range, i.e. very close to the robot and covering a large amount of its field of view, are represented as proto-objects. With these proto-object in its peri-personal range the robot does interact. The concept of peri-personal range reflects observations from the way small children perceive the world [14]. Additionally, the proto-object concept also covers visual stimuli in an inter-personal distance (here 1 - 2 m away). Their instantiation is solely based upon proximity, i.e. depth. They are not interacted with by the robot, but are used as top-down information for the auditory attention.

4.2. Auditory Attention

As a consequence of the long distance between the speaker and the microphones on the robot a large variety of signals overlay with the speech signal. For most robots the noise generated by the robot itself plays an important role. In our case this includes the noise generated by its arm and leg movement but also the noise emanating from its cooling fans mounted on its back, as head movements change the relative position of the microphones to the fans.

In a bottom-up stage the contrast enhancement between the environmental noise and the speech signal is mainly achieved by reducing the background noise based on beamforming techniques and adaptive noise level estimation.

Especially for instationary sounds additional top-down mechanisms are necessary. We investigated modulation based on the spectral characteristics of the speech and noise signals and an analysis of the motion status of the robot to suppress movement noise.

Another very important top-down information we recruit is the current interaction status of ASIMO which we determine based on the visual part of the attention system. When ASIMO neither sees an object in its peri-personal space or a human in its inter-personal space it assumes that nobody is interacting with it and hence it raises the minimal activity threshold for its auditory attention. This is a first step to suppress speech from people currently not interacting. In future we will replace this by better models of the interaction status of the tutor, e.g. based on gaze estimation.

5. Audio Visual Association Learning in Interaction

The design of our interactive learning system targets on bootstrapping multimodal representations with minimal initial knowledge and enabling a continuous development by learning in interaction. For instance our system can learn a cluster in the relative visual position space, an arbitrary speech label, and the association between both. We use some pretrained phrases which can trigger a learning session, e.g. "Learn where this object is.". A typical learning session consists of the following steps:

1. The tutor enters the interaction range of ASIMO so that it either sees the tutor or an object he is presenting.
2. The tutor utters one of the predefined learning phrases to teach categories as relative position, size, or a label to a movement of ASIMO.
3. The tutor presents an instance of the cluster to be learned, e.g. by showing and moving an object in the left field of view of ASIMO, while uttering the label he wants to associate to this cluster a few times (5-8).
4. When the tutor keeps silent for a few seconds the system ends the learning session and shows only reactive behavior.

To evaluate what the system has learned the tutor presents an object in one of the learned clusters and utters the associated label. If the active cluster and the recognized cluster do match ASIMO nods with its head. Otherwise ASIMO shakes its head and continues trying to find matches. If in a given time the match is found ASIMO finally nods and disables the expectation. The speech based interaction in this system is solely based on the microphones mounted on the robot and controlled via the attention system described in Sec. 4 [10]

Initially the system has only very little knowledge. The visual clusters and the speech labels are fully learned in interaction. During a learning session samples in the different perceptual modalities are accumulated. Within a session an object with the property to be labeled is presented, and matching speech labels are uttered several times. After a session has timed out, speech and the visual subsystem in focus determine the novelty of the current session to existing clusters. For each pair of two associated clusters a weighted summation of their activations is performed, forming a multimodal novelty signal. These signals are returned to their originating classifiers which individually decide whether the session data should be represented by a new cluster or whether the best matching cluster should be adapted. Finally, newly created clusters are associated with each other.

5.1. Online Word Learning

If no speech models have been learned yet a new model is initialized with the best matching phone sequence learned as described in Sec. 3. In later learning steps the current utterance is compared to the best matching speech label and the best matching phone sequence. If the novelty of the new label is strong a new cluster is learned, either based on the best matching cluster or the best matching phone sequence. If the novelty is weak the best matching cluster is updated. This allows the adaptation of already existing clusters.

5.2. Online Visual Cluster Learning

For learning of visual properties different features of the currently focused proto-object are used, such as a vector of its 3d position or the absolute value of its 3d size. Each cluster is represented by a multi-dimensional Gaussian, consisting of a cluster-center and a covariance. The activation of each cluster given some feature-vector is based on the distance between the cluster-center and the feature vector, integrated over time. The larger the distance, the lower the cluster activation [13]. In the same fashion as for the speech label learning it is also determined based on a novelty measure if a new cluster has to be created or rather an existing cluster should be updated.

5.3. Online Associtation Learning

Initially, the system neither contains any clusters nor associations. The learning of new associations assumes synchronously presented clusters in two different modalities to belong together. Therefore, the local learning decisions of the speech and the visual classifier in focus can be used to define the mapping between the two modalities. If both classifiers vote for the creation of a new cluster these two clusters are associated with each other. In the case where only one learning decision demands the creation of a new cluster this new cluster is associated with the already existing one in the other modality.

6. Speech Imitation

The communication of the robot with the tutor we described so far was solely based on the movements of the robot's body. To equip the robot also with verbal capabilities we investigate how sounds and words can be imitated based on the previously learned acoustical representations. As our system per se is not constrained in the vocal tract shapes it can model a direct imitation would result in a replica of the tutors voice. To avoid this we artificially impose such constraints to give it a child's voice. This entails the necessity to learn a mapping between the tutors voice and the system's voice. For the imitation this mapping is learned in interaction with the tutor. During synthesis motor primitives, manifest as vectors of formant positions in the child's voice space, are morphed to form a continuous speech segment. Such formant configurations can be also extracted in interaction from the tutor's voice [15, 16].

As a consequence the system can imitate utterances of the tutor with its own voice. At the current state this is limited to the imitation of vowel sequences. Interjacent consonants are filled with the best matching vowels [17].

7. Conclusion

Models of speech acquisition have to take into account different levels of abstraction and integrate information from modalities other than only sound. The acquisition of language relies on a shared realm of experience and knowledge between the child and its caregivers. Hence, in the development of models for speech acquisition we have to cover a wide range of topics to create this shared experiences and knowledge space between a robotic artifact and its tutor. The common theme behind the various aspects of our work we presented in this paper is the effort to integrate knowledge from developmental psychology and neurobiology into a model for speech acquisition embedded into a tutor-robot scenario.

8. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 28, no. 4, pp. 357–366, 1980.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans Speech and Audio Proc.*, vol. 2, no. 4, pp. 578–589, 1994.

[3] X. Domont, M. Heckmann, F. Joublin, and C. Goerick, "Hierarchical sectro-temporal features for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, Las Vegas, Nevada, 2008, pp. 4417–4420, IEEE.

[4] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *submitted to Speech Communication*.

[5] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex.," *Nat Neurosci*, vol. 6, no. 11, pp. 1216–1223, Nov 2003.

[6] N. Iwahashi, "Robots that learn language: Developmental approach to human-machine conversations," in *Lecture Notes in Computer Science*, vol. 4211, p. 143. Springer, 2006.

[7] Holger Brandl, Frank Joublin, Britta Wrede, and Christian Goerick, "A self-referential childlike model to acquire phones, syllables and words from acoustic speech," in *7th International Conference on Development and Learning*, Monterey, CA, USA, 10/08/2008 2008, IEEE, pp. 31–36, IEEE.

[8] M. Heckmann, H. Brandl, J. Schmuedderich, X. Domont, B. Bolder, I. Mikhailova, H. Janssen, M. Gienger, A. Bendig, T. Rodemann, M. Dunn, F. Joublin, and C. Goerick, "Teaching a humanoid robot: Headset-free speech interaction for audio-visual association learning," in *Proc. 18th IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*, Toyama, Japan, 2009, IEEE.

[9] B. Bolder, H. Brandl, M. Heracles, H. Janssen, I. Mikhailova, J. Schmuedderich, and C. Goerick, "Expectation-driven autonomous learning and interaction system," in *IEEE-RAS Int. Conf. on Humanoid Robots*. 2008, IEEE-RAS.

[10] M. Heckmann, H. Brandl, X. Domont, B. Bolder, F. Joublin, and C. Goerick, "An audio-visual attention system for online association learning," in *Proc. INTERSPEECH*, Brighton, UK, 2009, ISCA.

[11] J. B. Fritz, M. Elhilali, S. V. David, and S. A Shamma, "Auditory attention–focusing the searchlight on sound," *Current Opinion in Neurobiology*, vol. 17, no. 4, pp. 437 – 455, 2007, Sensory systems.

[12] L. Itti and C. Koch, "Computational modelling of visual attention," *NATURE REVIEWS NEUROSCIENCE*, vol. 2, no. 3, pp. 194–204, 2001.

[13] J. Schmuedderich, H. Brandl, B. Bolder, M. Heracles, H. Janssen, I. Mikhailova, and C. Goerick, "Organizing multimodal perception for autonomous learning and interactive systems," in *IEEE-RAS Int. Conf. on Humanoid Robots*. 2008, IEEE-RAS.

[14] C. Yu, L.B. Smith, and A. Pereira, "Grounding Word Learning in Multimodal Sensorimotor Interaction," in *Proc. of the 30th Annual Meeting of Cognitive Science Society (CogSci 2008)*, Washington DC, USA, 2007.

[15] C. Gläser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and bayesian estimation for robust formant tracking," *to appear in IEEE Trans. Audio, Speech, Lang. Process.*, 2009.

[16] M. Heckmann, C. Gläser, M. Vaz, T. Rodemann, F. Joublin, and C. Goerick, "Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS)*, Nice, 2008, IEEE-RSJ.

[17] M. Vaz, H. Brandl, F. Joublin, and C. Goerick, "Learning from a tutor: embodied speech acquisition and imitation learning," in *Proc. Int. Conf. on Development and Learning*, Shanghai, 2009, ICDL.