Learning from a tutor: embodied speech acquisition and imitation learning

Miguel Vaz, Holger Brandl, Frank Joublin, Christian Goerick

2009

Preprint:

This is an accepted article published in Proceedings of International Conference on Development and Learning. The final authenticated version is available online at: https://doi.org/[DOI not available]

Learning from a tutor: embodied speech acquisition and imitation learning

Miguel Vaz, Holger Brandl, Frank Joublin, and Christian Goerick

***Abstract*—This work presents a new developmentally inspired data-driven framework to bootstrap speech perception and imitation abilities in interaction with a tutor. The proposed system architecture extends our work presented in [1], that implements a cascade of interconnected layers to acquire the structure of speech in terms of phones, syllables and words. Here, we show how to couple such a perceptual model with a speech imitation system that produces speech sounds with a child's voice.**

The speech imitation system has at its core a correspondence model that links the tutor's and the system's voice. We present an interaction scheme for learning this correspondence model, where a human tutor provides the system with imitative feedback. Through this scheme, the system links its own vocal primitives with the tutor's voice. The correspondence model can then be used to map a tutor's utterance into a continuous activity in the system's motor space. These motor activities can then be used to imitate words.

Finally, we embed this architecture into an embodied autonomous learning and interaction system to provide a grounding for the speech models to be acquired and a perceptual input to trigger speech production.

***Index Terms*—speech imitation, speech acquisition, robotics, statistical language modeling**

I. Introduction

Social interaction between human and robot requires the robot to understand and to produce language. But these faculties are by no means trivial and need to develop in interaction with a caregiver. One important first element of speech and language understanding is the ability to parse spoken utterances into words. Another is mapping its perception of the tutor's speech to its own articulatory space to realize word imitation abilities.

Whereas linguistics and psychology have provided some insights about the way humans acquire the structure of language by a multitude of highly coupled bootstrapping processes, almost all speech processing systems neglected to provide a computational explanation for this complex learning process. Notable contributions to the problem of utterance parsing are [2], [3] and [4], that provided first insights into computational requirements and processes relevant for speech acquisition.

M. Vaz is with the Department of Industrial Electronics, University of Minho, Guimarães, Portugal e-mail: mvaz@dei.uminho.pt

H. Brandl is with the Research Institute for Cognition and Robotics, University of Bielefeld, Germany, email: hbrandl@techfak.uni-bielefeld.de

All authors are with the Honda Research Institute Europe, GmbH, Offenbach am Main, Germany

There is also a handful of models offering insights to the infant's acquisition of speech production skills and babbling processes, [5] [6]. Most of them, however, do not account for how infants address the correspondence problem: how to map acoustic targets proposed by the caregiver into some that are achievable by its different vocal tract. Recent work, [7] [8], has started transferring the burden of judging the phonological equivalence between caregiver's and infant's utterances from the infant to the caregiver, suggesting that the imitative response of the caregiver plays an important role in guiding infants to phonetically clear and meaningful utterances. This feedback is interpreted either as a reward [7] or as an unconscious correction signal [8].

Previously, we have presented a developmentally inspired purely data-driven model [1] for early infant word learning that attempts to acquire the structure of speech within a layered architecture comprising phone, phonotactics and syllable learning. Here, we extend this model with a scheme of how our robot learns to imitate its tutor using its own voice (figure 2) in a similar way to what we have already reported in [9]. The new model implements a tight coupling of perception and production, namely a correspondence model between phones acquired by the system and motor primitives innate to our robot. This coupling is learned through an exploratory process, in which the system learns the consequences of its vocal actions, in terms of the tutor's voice.

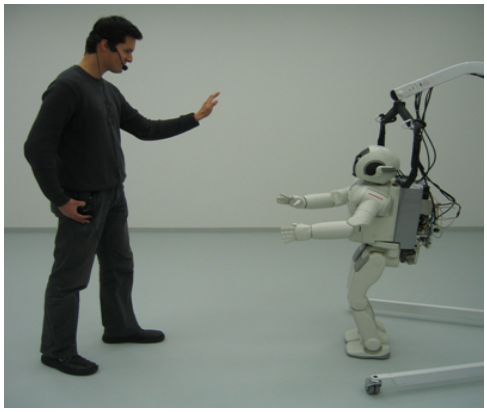

Fig. 1. Human tutor interacting with Honda's ASIMO. This interaction would be greatly improved by giving the robot the possibility of providing acoustic feedback about the scene or itself, using the vocabulary learned in interaction with the tutor.

Using statistical inference, our system converts a tutor utterance into a probabilistic sequence over the system's vocal

978-1-4244-4118-1/09/$25.00

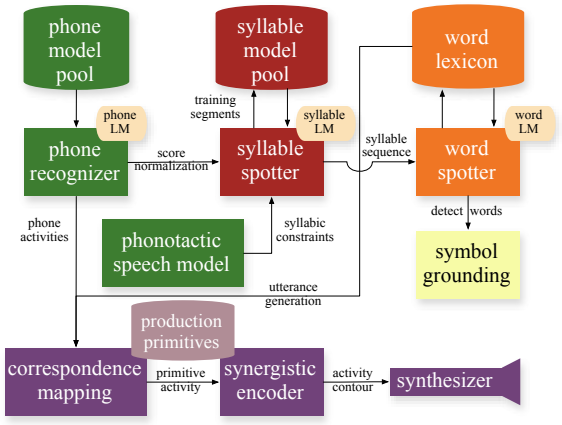

Fig. 2. The proposed architecture for coupled speech acquisition and production. All three acquisition layers have a very similar structure consisting of a pool of unit-models, a statistical grammar (LM), and a recognizer that detects learned units in the incoming feature stream. The coupling with the imitation layer takes place at the phone-level. A correspondence mapping transforms phone sequences into vocal primitives' probabilities, which are interpreted by the synergistic encoder as a continuous motor activation and sent to the speech synthesizer.

repertoire, that is subsequently transformed into a synergistic motor coding, used to imitate the tutor utterance. To evaluate our integrated speech acquisition and production model, we present an interaction experiment between a human tutor and our robot.

This work is outlined as follows. We introduce all parts of our system in II, comprising the speech acquisition module, the correspondence model, the synergistic encoder, and the synthesis module. In section III we describe the details of the used interaction schemes between human and robot. We present the results of our experiments in section IV, and conclude with a short discussion in V.

II. System Architecture

Our system architecture is driven by our previous research towards embodied autonomous learning and interaction as described in [10]. There we presented a system to associate object properties and actions to auditory labels in interaction with our humanoid robot ASIMO. Here, we aim to equip it with the ability to describe a set of perceived object properties (like size, position, motion state) to associated auditory labels previously learned in interaction, that can be subsequently imitated with a child's voice. We focus on developing not only a technical framework, but rather a (simplified) model of early infant word learning and production. Therefore, we investigate how to couple our developmentally inspired framework for speech acquisition proposed in [1] with a speech mimicking system started in [9].

A. Three-layered acquisition architecture

Most computational models for word acquisition suffer from two major weaknesses. First, they tackle the problem of speech acquisition in the symbolic domain only. However, it is not clear how and whether these approaches can be generalized to the acoustic domain. Second, most models rely on some kind of innate representation, which is mostly at the level of syllables. But since syllables strongly depend on the language to be learned, it is not clear how these approaches can be extended to become valid models for speech acquisition as observed in infants.

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. [11], [4]), we proposed a three-layered framework for speech acquisition in [1]. The idea of this work was to bootstrap a word representation incrementally based on the statistics of raw acoustic input speech only. It is implemented as a cascade of three HMM-based speech unit spotting instances that rely on incomplete speech unit representations on phone, syllable and word level. Each layer comprises a pool of speech unit models, a detector and a statistical speech unit grammar that is estimated from the layer's recognition results.

Its first layer bootstraps a phone-representation and provides phone recognition and segmentation results for subsequent processing steps. Inspired by [2], Mel-frequency cepstrum coefficients of a few minutes of input speech are accumulated, providing the system with a sufficiently large training sample. Single state HMMs with mixtures of Gaussians, including 8 component densities as output probability distribution functions (OPDF), are estimated using k-means clustering. The actual phone models are created from most frequent state-sequences, as obtained by a Monte-Carlo-sampling between these single states embedded into an ergodic HMM. These state sequences are concatenated to 3-state phone-models with Bakis-topology and become further refined using Baum-Welch training.

To allow the learning of syllables, phone recognition results are condensed into a *phonotactic model*, which aims to describe the probabilistic phone structure of syllables in the tutor's language. This completely priors syllable learning, because it is the only sufficiently reliable linguistic cue for parsing utterances into syllables. Especially length of utterances is not a reliable cue to decide what a syllable is. It is technically implemented as a pair of Katz-smoothed trigram models on phone symbol level to encode initial and final parts of the syllabic structure. Both are estimated from phone results from the initial and final parts of the input utterances, because these are the only syllable boundaries that can be reliably detected without assuming an innate syllable parser.

This phonotactic model provides a parsing into syllable segments that found the basis for the second layer as shown in figure 2. This layer implements an incremental clustering scheme on syllable segment level, that bootstraps a syllable representation. This involves a novelty detection step and a subsequent model update or the creation of a new syllable model in case that a segment does not match sufficiently to any existing syllable model. Initially the syllable representation does not contain any models.

Finally, our framework acquires a word lexicon using a

bootstrapping scheme that combines the principle of subtraction [11], the finding that child directed speech is dominated by short utterances, and statistical learning [4]. Because we assume speech variability to be mainly bound to the level of syllables, words models abstract from the acoustic domain and are represented as syllable symbols sequences: Instead of acoustic used as input for phone and syllable layer, the sequence of spotted syllables defines the sole input to lexical acquisition and word recognition (c.f. [1] for details).

B. Speech imitation layer

Finding own motor configurations that produce phonetical equivalents of tutor's words is crucial to speech imitation and acquisition. This is called the correspondence problem and is by no means a trivial task to solve, due to significant differences between the voices of the caregiver and the infant. Different lengths and proportions of their vocal tracts cause these differences, which include higher pitch and formant frequencies. Yet children are able to progressively solve this problem, already showing regional-dependent phonetic preferences in their early babbling process and the ability to produce simple words without any trace of foreign accent around the age of 1 [12].

Some speech acquisition models do not account for it [13] [5], while others [12] argue that this ability is innate. However, a set of acoustic features where phonetically similar speech sounds share the same representations is yet to be found. This raises the question as to whether children use other information to address the correspondence problem. Recent work indicates that extra knowledge is unconsciously offered by the tutor, while interacting and imitating with the child: [8] shows how the phonological bias of the tutor can guide the infant to clearer vowel sounds during a mutual imitation game, and [7] suggests a reinforcement learning interpretation of the acquisition of vocal skills, under which parental feedback constitutes a reward signal, and where increasingly demanding parental standards constitutes the necessary force for the infant to improve its vocal speech skills.

Recently [9], we have shown how a correspondence can be found between vowel sounds from a system with a child's voice and the equivalent vowels from its tutor, making no assumptions on the language of interaction or the phonetical properties of the tutor or system's voice, and relying solely on an imitative response of the tutor. We hereby adapt this previous work to integrate with the aforementioned speech acquisition system. As shown in figure 2, the new imitation subsystem consists of three main modules: a probabilistic mapping between the phone models acquired by the system and its own vocal repertoire, a synergistic encoder that converts a probabilistic distribution over the set of motor primitives into a motor activation, and a synthesis module that synthesizes motor activation into speech.

We will make use of the following notation throughout this work: λ_i^p represents phone model i, $[\lambda^p]$ a sequence of phone models, $\mathcal{P}$ the set of all possible phone sequences, X_{tutor} an acoustic observation from the tutor, m_j stands for motor primitive j. The terms *vocal* and *motor primitive* will be used interchangeably.

C. Probabilistic phone correspondence

The probabilistic phone mapping $\mathcal{C}$ here presented is learned using a similar interpretation of the role of the tutor feedback as the one described in [9]. The difference lies in the tutor feedback being now a sequence of recognized phone models, instead of acoustic features. This represents a significant improvement in several aspects. One of them is robustness, because phone models are trained with better perceptual features and their classification is performed using a state-of-the-art method. The model is also more plausible, because it provides a tight coupling between perception and production. Furthermore, interaction with the system is also more natural, because both perceptual and production skills can be trained in a unified scenario, without hard boundaries between training and testing phases.

The correspondence mapping $\mathcal{C}$ represents, for each phone model λ_i^p, a probability distribution over the space of motor primitives.

$$\mathcal{C}_{ij} = P(m_j|\lambda_i^p) \tag{1}$$

In section III-B we describe how this mapping is learned.

D. Synergistic encoder

Given an utterance to be imitated, the synergistic encoder computes a vocal output from a sequence of recognized phone model hypotheses, provided by the phone recognizer (c.f. figure 2). Each segment hypothesis λ_i^p has an associated time span ($[t_0, t_1]$) and a probability distribution over the vocal primitives, given by the correspondence mapping $\mathcal{C}$. The synergistic encoder uses these probabilities to compute an activation curve for each vocal primitive, which takes the form of a gaussian with mean in the middle of the activation interval, standard deviation proportional to its width, and magnitude given by the correspondence mapping C_{ij}. To compute the overall activation of each motor primitive, $\mathcal{W}_{m_j}$, we sum the activation contours from all segment hypotheses:

$$\mathcal{W}_{m_j}(t) = \sum_{\lambda_i^p} \mathcal{C}_{ij}\, G\Big(t\,;\, \mu = \frac{t_1+t_0}{2},\, \sigma = k\,(t_1 - t_0)\Big) \tag{2}$$

The reason behind the choice of a gaussian form for the individual activation contours is that it enables the simulation of some coarticulation effects, namely smooth transitions between consecutive segments.

E. Synthesis module

For the synthesis of speech, we use a vocoder-like scheme developed for equipping the system with the possibility of synthesizing high-pitched voices, like children's. In spite of recent developments, articulatory models do not yet offer this possibility. Also, traditional speech coding techniques show limitations with high-pitched voices, although recent work has shown child speech with good quality [14].

This algorithm, described in more detail in [9], makes use of a gammatone filter bank at its core, which allows for an optimal tradeoff between spectral and temporal resolution. As a consequence, high and low pitched voices can be synthesized with similar quality and without the need of any special speaker-dependent model.

For this synthesizer, motor primitives have the form of spectral vectors, annotated with their formant frequencies for use of morphing algorithm. The set of motor primitives and their activations is transformed into a unique spectral output with a speech morphing algorithm, and sent to the synthesizer.

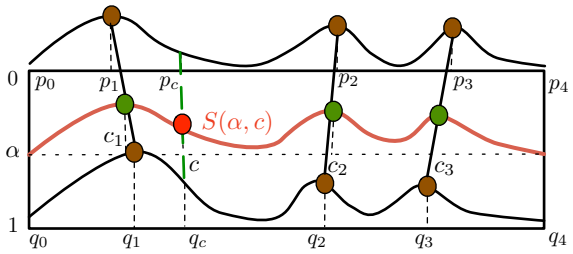

Fig. 3. The morphing algorithm computes the spectral value $S(\alpha, c)$ for a channel c at an intermediary position α given an initial and final spectral vectors p and q, and a correspondence matrix associating (p_i, q_i).

1) Spectral morphing algorithm: For a target utterance, there will probably be more than one vocal primitive m_j with a positive activation at a given time instant, $\mathcal{W}_{m_j}(t)$, as defined by equation 2 It is therefore necessary to transform the activate vocal primitives into a single spectral output to be fed to the synthesis algorithm. This is accomplished by means of a morphing algorithm, $\mathcal{M}$. Morphing two spectral vectors, m_j and m_k, results in a third spectral vector representing an intermediate state, where the value of each spectral channel is given by

$$\begin{aligned} \mathcal{M}(m_j, m_k, \alpha_j, c) &= (1-\alpha_j)\, m_j(p_c) + \alpha\, m_k(q_c) \\ \alpha_j &= \frac{\mathcal{W}_{m_j}}{\mathcal{W}_{m_j} + \mathcal{W}_{m_k}} \end{aligned} \tag{3}$$

Here, $m_j(c)$ refers to channel c of m_j, and p_c and q_c are calculated by maintaining the proportion of the distance from channel c to the immediately inferior q_c, respectively part of the initial and final spectral vectors. recursively morphing each of the motor primitives, as shown in figure II-E.

III. Interaction

In [10] 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. Our motivation since then has been to extend it with the ability to describe a presented object. For example, when presented with a red apple on the right side, our system should be able to provide an acoustic scene description like "right red apple".

For this, the system needs to be able to project the acoustic targets of the learned labels into its own articulatory space. In our system, this correspondence model $\mathcal{C}$ is encoded on the a phonemic level, and is learned through interaction with the tutor. We integrated this learning in the overall interaction scheme, by making our system to initiate interaction after a given period of inactivity: The system produces one of its basic vocalic sounds, and uses the tutor's imitative response to train the probabilistic correspondence mapping.

A. Tutor imitates system

In the training phase the system learns a correspondence between its motor primitives and imitative responses from the tutor. It produces vowel utterances with constant timbre by synthesizing spectral vectors from its repertoire. The tutor then imitates the system, which determines the best matching phone sequence

$$[\lambda_1^p, ..., \lambda_n^p] = \arg \max_{[\lambda^p] \in \mathcal{P}} P([\lambda^p] | X_{tutor}) \tag{4}$$

The experience mapping M representing the probability of perceiving phone model λ_i^p given a vocal primitive m_j

$$M_{ij} = P(\lambda_i^p | m_j, Dj) \tag{5}$$

is then updated in proportion to the segment length of each detected phone model λ_i^p. This procedure is schematized in figure 4.

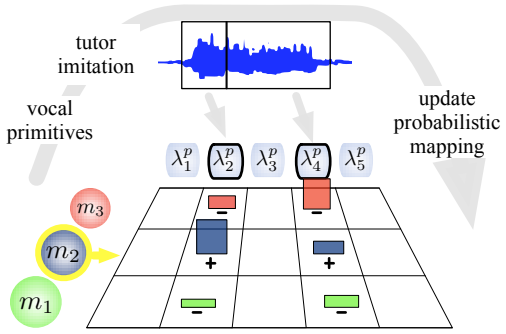

Fig. 4. Example for correspondence model learning. A randomly picked vocal primitive m_2 is synthesized with constant timbre. The tutor imitates the vowel sound and the response is used to update the experience mapping in proportion to amount of activation of each phone.

B. System imitates the tutor

In order to imitate, the system maps phone model likelihoods to activations of vocal actions, using the probabilistic correspondence mapping described in equation 2.

The correspondence mapping is inferred from the experience mapping, see equation 5.

$$C_{ij} = P(m_j | \lambda_i^p) = \frac{P(\lambda_i^p | m_j, Dj)\, P(\lambda_i^p)}{P(m_j)} \tag{6}$$

Because we assume a flat prior over all motor primitives, and values of the mapping are only computed considering a single phone model at a time, the correspondence mapping can be represented as

$$C_{ij} = M_{ij} \tag{7}$$

The likelihood of each motor primitive is passed onto the synergistic encoder, which computes a time sequence of motor primitive activations as described in section II-D. This sequence is then recursively morphed into a single vocal output for each time instant, according to the motor primitives' relative strength of activation, and passed onto the synthesis algorithm that generates the imitation signal.

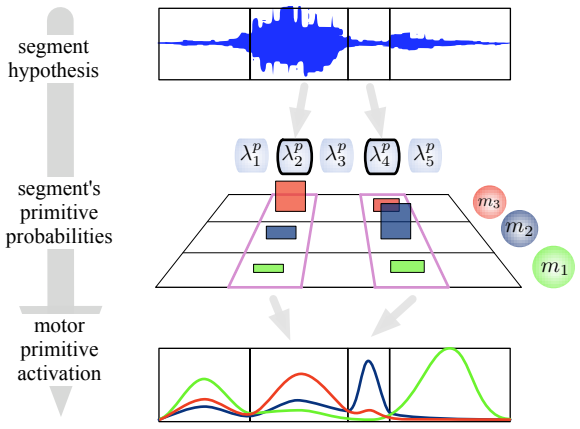

Fig. 5. A tutor utterance to be imitated is parsed into a sequence phone segments. Using the correspondence mapping, the probability of each motor primitive for the phone models most active in the different segments is computed and used by the synergistic encoder to generate the motor activation.

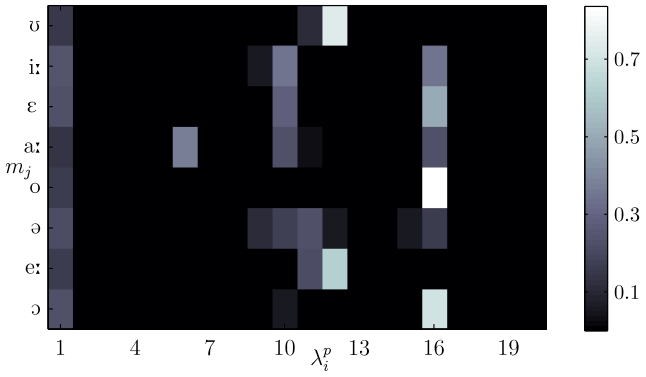

Fig. 6. An instance of a correspondence model learned in interaction with a tutor. The phone OPFs conditioned with the different motor primitives are shown in the rows. The vowel phones (represented using IPA notation) are marked.

IV. Experimental results

Given a learned phone representation, we evaluated the correspondence model bootstrapping as described in section III-A. We grounded the system's voice in a set of 8 spectral vectors, selected from cluster centers computed with the K-Means algorithm over the spectrograms of utterances spoken by a 10 year old male child, from the TIDIGITS corpus [15]. Each spectral vector, 100 channels with center frequencies from 40Hz to 8KHz, represents one of the following vowels (IPA alphabet): ɔ, e, ə, o, a, ɛ, i, ʊ.

Each of the vocal primitives was synthesized and played 15 times to a male adult speaker, who imitated them (cf. sec. II-C). We synthesized each robot's utterance with a random duration (between 0.25 to 0.3 seconds), and different pitch contours. The resulting correspondence model can be seen in figure 6.

The following aspects can be observed from the data. Firstly the imitative response of the tutor only covers a subset of the set of phone models. This was expected, because the system is limited to the production of vowel sounds, and the phone models are trained using unconstrained speech containing both vowels and consonants.

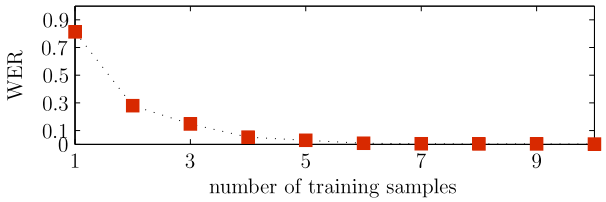

Fig. 7. Mean word classification performance with respect to the number of training samples

Secondly, phone model with index 1 has a very strong response for all tutor responses; this is an artifact due to our voice activity detection that includes short noise parts in the beginning and at the end of the detected speech segments.

Thirdly, the models for the different vocal primitives vary considerably: primitives for vowels ɔ, e, ʊ have a very unimodal response, while others like ɐ have a more disperse response. Several factors might be contributing to this, the most likely being either a non-uniform imitative response of the tutor to the vocal primitive or the inexistence of any phone model fully representing the imitative response. One reason supporting the first might be that, although the vocal primitives were selected with care to correspond to one vowel, synthesizing a sound with constant timbre presents limitations to its naturalness, not necessarily affecting all vowel sounds equally. One reason supporting the second, is that the phone models are trained using different data, even if originating from the same speaker. We tried to compensate this effect by balancing the words in the training corpus according to the vowels contained, but issues with over- or under-representation are seldom avoided in unsupervised learning systems. Another possibility would be to (at least partially) overlap the phone model learning phase with the learning of the correspondence model, so that the phone models can be estimated using similar data. The disadvantage would be that the interaction phase would take longer.

An example of the different stages of processing can be seen in figure 8, where the word *mama* is imitated. As already explained, only the vowel segments are being imitated. Thus, the words the system produces can be distinguished if the vowel constituent's sequence is different.

Additionally, we also evaluated the learning of new words, performed as follows. Given an object, the tutor focuses the system attention to an object property that should be labeled (e.g. size or relative position to the robot) (see [10] for details). The tutor then provides a few (2-5) isolated samples for each word. The temporal grouping of these speech segments is given to the system as an additional cue, in order to ease learning. The system was presented with a total of 20 different words, predominantly mono- or bisyllabic, like *red*, *green*, *bottle*, *duck*, among others. As expected, the more the speech segments provided in a learning session, the better the classification results (as shown in figure 7)

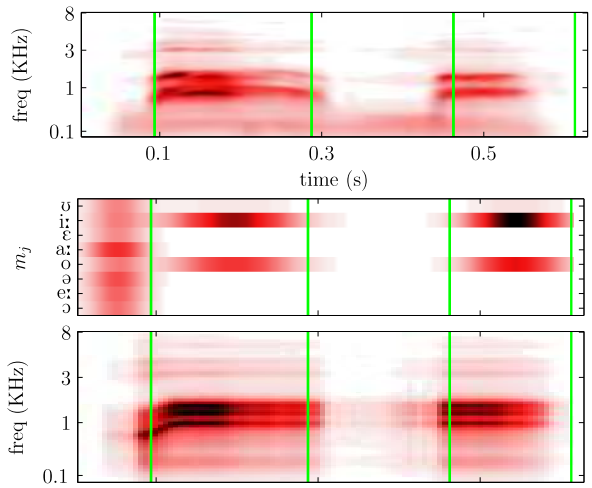

Fig. 8. An example of an utterance imitation performed by our system. For an input utterance (enhanced spectrogram in the upper figure), the most likely phone model sequence is determined, dividing the utterance in different segments (delimited by the green vertical lines). The posterior probability of the motor primitives for each phone hypothesis in the sequence is computed, and used as a strength of activation (middle figure). The output spectrum, used for the synthesis, is shown in the bottom figure.

V. Conclusion

We presented and tested an integrated approach for infant-inspired speech acquisition and production by coupling an embodied data-driven perception system with an imitation system. By assuming a cooperative tutor that imitates monophonic utterances of the system, we've shown how to bootstrap a probabilistic correspondence model of the tutor's imitative response to each of the system's motor primitives. We've also presented how such a model equips our humanoid robot with the ability to describe its environment in terms of labels for various object properties that have been associated with arbitrary words in interaction with a tutor.

Our approach extends previous attempts [9] for sensory-motor coupling, because it involves more and more plausible training data to estimate the perceptual part of the system: the phone models are estimated not only from the examples collected during the imitation learning, but also from the whole history of interaction. This way, the models are learned not only from isolated phone-instances from the tutor-imitation, but also from a continuous speech context. Although the vocal repertoire of the system contains only vowels, which obviously impairs the complete imitation of words containing consonantal sounds, we consider this to be an important step towards embodied online learning of speech and language abilities.

Our next steps are three-fold. Firstly, we plan to investigate an even tighter coupling of production and perception. This includes a more elaborate utterance generation, in the case of several words, and the possibility of learning production models of perceptual syllable models in interaction with a caregiver. For bootstrapping this process, the phone-activation signal could be used, and then changed through interaction.

Secondly, we would like to add consonant sounds to the system's repertoire; namely steady-state consonants like fricatives or nasals could be learned using a similar interaction scheme as the one hereby used for vowel sounds.

Thirdly, it would also be interesting to ease the strict assumption of the cooperative tutor, even though the frequentist nature of the probabilistic correspondence model gives it some robustness. Namely, that the system can discriminate between imitative and non-imitative tutor responses.

Acknowledgment

We'd like to thank Dr.Estela Bicho and Dr. Wolfram Erlhagen for their support. This work was supported by a doctoral grant (SFRH-BD-12637-2003) from the Portuguese Science and Technology Foundation, *FCT*. Part of this work by a scholarship from the Calouste Gulbenkian Foundation.

References

[1] H. Brandl, B. Wrede, F. Joublin, and C. Goerick, "A self-referential childlike model to acquire phones, syllables and words from acoustic speech," *7th IEEE Int. Conf. on Development and Learning*, pp. 31–36, 2008.

[2] N. Iwahashi, "Robots that learn language: Developmental approach to human-machine conversations," in *Symbol Grounding and Beyond - EELC*, P. Vogt, Y. Sugita, E. Tuci, and C. Nehaniv, Eds., 2006, pp. 143–167.

[3] K. Gold and B. Scassellati, "Audio speech segmentation without language-specific knowledge," in *Cognitive Science*, 2006, pp. 1370–1375.

[4] R. N. Aslin, J. R. Saffran, and E. L. Newport, "Computation of conditional probability statistics by 8-month-old infants," *Psychological Science*, vol. 9, no. 4, pp. 321–324, July 1998.

[5] F. Guenther and J. Perkell, "A neural model of speech production and its application to studies of the role of auditory feedback in speech," in *Speech Motor Control in Normal and Disordered Speech*, B. Maassen, R. Kent, H. Peters, P. V. Lieshout, and W. Hulstijn, Eds. Oxford University Press, 2004, pp. 29–49.

[6] G. Westermann and E. R. Miranda, "A new model of sensorimotor coupling in the development of speech." *Brain and language*, vol. 89, no. 2, pp. 393–400, Apr 2004.

[7] I. S. Howard and P. R. Messum, "A computational model of infant speech development," *XII Int. Conf. Speech and Computer - SPECOM*, 2007.

[8] K. Miura, Y. Yoshikawa, and M. Asada, "Unconscious anchoring in maternal imitation that helps finding the correspondence of caregiver's vowel categories," *Advanced Robotics*, vol. 21, pp. 1583–1600, 2007.

[9] M. Vaz, H. Brandl, F. Joublin, and C. Goerick, "Speech imitation with a child's voice: addressing the correspondence problem," *accepted for 13-th Int. Conf. on Speech and Computer - SPECOM*, 2009.

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

[11] P. W. Jusczyk, "How infants begin to extract words from speech," *Trends in Cognitive Sciences*, vol. 3, no. 9, pp. 323–328, September 1999.

[12] P. K. Kuhl and A. N. Meltzoff, "Infant vocalizations in response to speech: vocal imitation and developmental change," *Journal of the Acoustical Society of America*, vol. 100, no. 4 Pt 1, pp. 2425–38, 1996.

[13] B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, September 2008, in press.

[14] O. Watts, J. Yamagishi, K. Berkling, and S. King, "HMM-based synthesis of child speech," in *Proc. of The 1st Workshop on Child, Computer and Interaction (ICMI'08 post-conference workshop)*, Crete, Greece, October 2008.

[15] R. G. Leonard, "A database for speaker-independent digit recognition," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP*, vol. 9, pp. 328–331, 1984.