Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model

Anja Kristina Philippsen
Cognitive Interaction Technology Center (CITEC), Bielefeld University
anja.philippsen@uni-bielefeld.de

René Felix Reinhart
Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University
freinhart@uni-bielefeld.de

Britta Wrede
Applied Informatics Group, Bielefeld University
bwrede@techfak.uni-bielefeld.de

***Abstract*—This paper proposes an efficient neural network model for learning the articulatory-acoustic forward and inverse mapping of consonant-vowel sequences including coarticulation effects. It is shown that the learned models can generalize vowels as well as consonants to other contexts and that the need for supervised training examples can be reduced by refining initial forward and inverse models using acoustic examples only. The models are initially trained on smaller sets of examples and then improved by presenting auditory goals that are imitated. The acoustic outcomes of the imitations together with the executed actions provide new training pairs. It is shown that this unsupervised and imitation-based refinement significantly decreases the error of the forward as well as the inverse model. Using a state-of-the-art articulatory speech synthesizer, our approach allows to reproduce the acoustics from learned articulatory trajectories, i.e. we can listen to the results and rate their quality by error measures and perception.**

I. Introduction

Speech production and the imitation of perceived sounds requires knowledge about how to control the articulators of the vocal tract in order to achieve desired acoustics. Knowledge of two mappings is required in this context: The forward mapping estimates which acoustics will result from a specific articulatory movement. This corresponds to the learner's expectation of which acoustics his vocal tract will produce in response to a motor command. The inverse mapping, in contrast, estimates which vocal tract movements are required in order to reproduce an acoustic signal. Knowledge of the inverse mapping enables acoustic imitation.

Especially the inverse mapping is extensively studied due to evidence that articulatory parameters are beneficial for speech recognition [1]. Proposed models to solve this non-linear and non-unique mapping include neural networks [2], statistical methods [3] and codebook approaches [4] and generally rely on large data bases of human recorded articulatory-acoustic data, e.g. the MOCHA data base [5]. However, such supervised learning does not explain how an agent can learn and refine its model for speech recognition and production. Developmental approaches which aim at modeling the autonomous acquisition of speech are promising, but often focus on rather restricted speech production skills, e.g. vowel production and reinforcement using formant space representations [6], [7], [8], [9], [10]. The biologically inspired DIVA model [11] and similar models (e.g. [12]) can produce consonants, but do not learn the inverse mapping directly. Instead acoustics and articulation are connected by a map of speech sounds. A recent related work is [13] where goal babbling is shown to lead to the emergence of consonant-like structures. But due to the low-dimensional acoustic representation their model does not distinguish between different consonants.

This paper tackles the question of how to efficiently learn the forward and inverse mapping for syllable sequences from few supervised training examples and how such initial models can be refined in an unsupervised manner by trying to imitate acoustic stimuli.

In a first step, we apply an efficient recurrent neural network approach to learn the forward and inverse model of speech production for syllable sequences which cover the coarticulation of eight vowels and eight consonants. The recurrent neural network model handles syllable sequences as continuous trajectories in the acoustic as well as in the articulatory space. The network dynamics account for temporal dependencies in the sequences.

In a second step, we train initial forward and inverse models on a small set of articulatory-acoustic example trajectories. These initial models are improved by presenting auditory goals that the learner tries to imitate. The acoustic outcomes of the imitations together with the executed actions serve as new training pairs for sequential learning. This refinement process is similar to goal babbling as implemented in [10] and [13], and unsupervised, i.e. it requires only acoustic data.

While the initial models can produce the vowels and consonants in specific contexts, this paper shows that it is indeed feasible to improve the generalization accuracy of such initial models in novel contexts by trying to imitate acoustic stimuli. This result contributes to earlier demonstra-

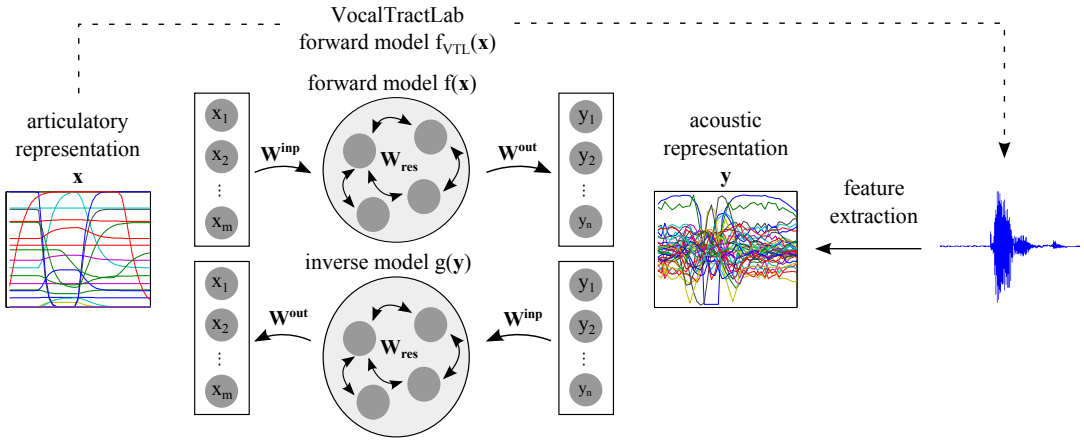

Fig. 1. The forward and inverse mapping between articulatory and acoustic parameter spaces.

tions of refinement, e.g. learning direct inverse kinematics of a robotic arm [14], and extends these results to the imitation of complex articulatory-acoustic feature trajectories with intricate temporal dynamics. In contrast to [14], the proposed method does not require the derivation of the forward or inverse model.

A learner using its own prediction to teach itself is also a common technique in the field of semi-supervised methods for pattern classification and known as self-training (e.g. [15]). In self-training, a confidence criterion is typically used to decide which unlabeled examples should be used for retraining with the estimated class labels. Here, instead the learner has the forward model (i.e. its vocal tract) available. This paper contributes to the research on self-training by showing that this principle can also be applied to learning of high-dimensional and continuous sequence transduction tasks.

In comparison to autonomous exploration techniques, e.g. using goal babbling ([16], [10]), learning in this paper is guided by a teacher similar to [6], [7], [9] and [13]. In [13], mappings were implemented as Gaussian mixture models, while we apply recurrent neural networks. Besides, the focus in [13] is on modeling the development of vocalization in terms of active exploration; utterances are characterized by formant values and intensity only. This work, in contrast, focuses on teaching the learner to reproduce a large set of different consonant and vowel sounds by using a rich acoustic representation.

We show that the error of the forward and inverse model can be decreased significantly by sequentially retraining both models with the estimated articulation and the corresponding acoustic outcome. The results are systematically evaluated by standard error measures and by perceptual tests, i.e. rating the acoustics produced by the model.

II. Learning the Articulatory-Acoustic Mapping

In this section a set of articulatory-acoustic sequences is introduced and a cross-validation test is conducted which verifies that mappings between the articulatory and acoustic representations of these sequences can be learned with excellent generalization errors by an efficient recurrent neural network model.

In the following, we refer to the model that maps from articulatory to acoustic space as the forward model f, while the inverse model g realizes the mapping from the acoustic to the articulatory space. Trajectories in the articulatory space are denoted by $\mathbf{x}(k)$ and acoustic trajectories by $\mathbf{y}(k)$, where k denotes the time step. We aim at training an artificial neural network to approximate the forward model $\mathbf{y}(k) = f(\mathbf{x}(k))$ as well as the inverse model $\mathbf{x}(k) = g(\mathbf{y}(k))$. To account for the dynamical properties of the feature trajectories, a recurrent neural network model is used to learn the forward and inverse model, respectively (cf. Fig. 1).

A. Articulatory-Acoustic Data Set

For data generation and evaluation of the estimated articulatory sequences, the speech synthesis system *VocalTractLab* developed by Birkholz [17] was used to generate acoustic signals from articulatory parameter trajectories. The articulatory parameters describe the positions of important articulators and the vocal tract shape (e.g. the tongue tip position, lip distance and jaw opening angle) as a function of time. The sequences were created manually using the phone definitions and the gestural scores representation provided by *VocalTractLab 1.0*. 22 out of the 25 tract parameters and 4 glottis parameters were used for the articulatory representation. [1]

The data set consists of 64 different articulatory sequences which start with the vowel [aː] and have the form <aCV>, where C is one out of eight consonants ([b], [d], [g], [z], [p], [t], [k], [ʃ]) and V is one out of eight vowels ([aː], [eː], [iː], [oː], [uː], [æː], [œː], [yː]). Each utterance is 500 ms long. Note that the data comprise four voiced and four voiceless consonants. Three of the voiced consonants are plosives and used together with their voiceless counterparts. The other two consonants are fricatives. For each of the 64 sequences, 50 noisy samples were generated by varying the consonant and vowel durations, the articulatory effort,

[1] Velum position and tongue center radius were omitted as they do not change within the data set.

and by adding noise to the lung pressure parameter and the fundamental frequency.

VocalTractLab generates acoustic signals based on the articulatory trajectories. As acoustic features we chose Mel-frequency cepstral coefficients (MFCCs), the standard features for speech recognition. The 39-dimensional feature vector contains logarithmic energy and the first 12 MFCCs as well as the first and second derivatives. All articulatory and acoustic sequences have been normalized to the range $[-1, 1]$.

B. Echo State Network Learner

For efficient learning of the forward and inverse models, we apply the so-called reservoir computing approach which separates a non-adaptive reservoir of recurrently connected neurons from an adaptive linear read-out layer. We apply a particular flavor of reservoir computing known as Echo State Network (ESN, [18]). ESNs comprise three layers of neurons: The input layer $\mathbf{u} \in \mathbb{R}^D$, a hidden layer of reservoir neurons $\mathbf{h} \in \mathbb{R}^H$, and the output layer $\mathbf{v} \in \mathbb{R}^O$. The reservoir state $\mathbf{h}$ is updated according to

$$\mathbf{h}(k) = tanh(\mathbf{W}^{\text{inp}}\mathbf{u}(k) + \mathbf{W}^{\text{res}}\mathbf{h}(k-1)),$$

where k is the time step of the discrete dynamics and $\mathbf{u}(k)$ the current input. The connection weights $\mathbf{W}^{\text{inp}} \in \mathbb{R}^{H\times D}$ from the input layer to the reservoir as well as the recurrent connections $\mathbf{W}^{\text{res}} \in \mathbb{R}^{H\times H}$ are initialized randomly and remain fixed. To assure that perturbations of the reservoir state by the input decay over time, i.e. cause an *echo* in the reservoir dynamics, the spectral radius of the reservoir matrix $\mathbf{W}^{\text{res}}$ is scaled close to 1 (cf. [18]).

Supervised training is restricted to the read-out weights $\mathbf{W}^{\text{out}} \in \mathbb{R}^{O\times H}$ which linearly combine the reservoir state to compute the output:

$$\mathbf{v}(k) = \mathbf{W}^{\text{out}}\mathbf{h}(k)$$

Given a set of input and target output pairs $\{(\mathbf{u}(k), \mathbf{t}(k))\}_{k=1,\ldots,K}$, training can be accomplished by linear regression according to

$$\mathbf{W}^{\text{out}} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbb{I})^{-1}\mathbf{H}^T\mathbf{T}, \tag{1}$$

where the matrices $\mathbf{H} \in \mathbb{R}^{K\times H}$ and $\mathbf{T} \in \mathbb{R}^{K\times O}$ row-wise collect the reservoir state and target outputs, respectively. The model complexity is controlled by the number of hidden neurons H and the regularization parameter $\lambda \geq 0$ in (1). Initial transients of the reservoir dynamics are washed out by feeding the network with the first vector of the new sequence for a number of time steps.

Note that inputs $\mathbf{u}$ and outputs $\mathbf{v}$ of the ESN take the role of articulatory parameters $\mathbf{x}$ and acoustic representation $\mathbf{y}$ for the forward model (cf. Fig. 1). Accordingly for the inverse model, inputs $\mathbf{u}$ correspond to $\mathbf{y}$ and outputs $\mathbf{v}$ to $\mathbf{x}$.

C. Forward and Inverse Mapping Results

We conduct a leave-one-sequence-out cross-validation test in order to assess the generalization performance of the ESN for the forward and inverse model. That is, we train ESNs on 63 out of the 64 sequences (including the 50 variations per sequence) and compute the generalization error on the left out sequence. The errors of the forward and the inverse models are both calculated in the acoustic space using the dimension-normalized Mean Square Error (MSE)

$$e = \frac{1}{K}\sum_k (\frac{1}{D}\sum_d ||\mathbf{y}(k)^d - \hat{\mathbf{y}}(k)^d||^2), \tag{2}$$

where $\mathbf{y}$ is the acoustic target trajectory and $\hat{\mathbf{y}}$ the model estimate. K refers to the number of discrete time steps and D is the dimensionality of the acoustic feature vector. The error of the forward model is calculated directly between the estimated and the target trajectories. For the inverse model, the estimated articulatory trajectories are transformed back to the acoustic space via *VocalTractLab* and then compared with the acoustic target sequences. We trained networks with $H = 300$ hidden neurons and regularization parameters $\lambda = 10^{-3}$ for the forward model and $\lambda = 10^{-6}$ for the inverse model. The connection weights in $\mathbf{W}^{\text{inp}}$ and $\mathbf{W}^{\text{res}}$ are drawn from a uniform distribution in $[-1, 1]$. The spectral radius of $\mathbf{W}^{\text{res}}$ is scaled to 0.95, the input weights are scaled with $1/D$.

The upper two rows in Tab. I show the results for the forward and inverse model, averaged over 10 repetitions of the cross-validation to account for the random network initialization. The results show that ESNs are capable to learn the forward and inverse model of the articulatory-acoustic mapping with low errors. Generalization errors are in the same range as the training errors which indicates proper generalization for both models. The error of the inverse model is slightly higher than the error of the forward model. This is mainly due to the special characteristic of the inverse mapping: It requires the production of smooth articulatory trajectories from rather jerky acoustic feature trajectories.

In addition to the results for the ESN with a dynamical reservoir (upper two rows in Tab. I), we also include results for a non-dynamic variant of the ESN without recurrent connections ($\mathbf{W}^{\text{res}} = \mathbf{0}$) which is known as Extreme Learning Machine [19] (ELM, lower two rows in Tab. I). Using the non-dynamic ELM, a higher error in the forward model can be observed. This indicates that the reservoir dynamics support the generation of distinct features of the jerky acoustic trajectories from the rather smooth articulatory sequences.

Tab. I
TRAINING AND GENERALIZATION MEAN SQUARE ERRORS OF THE FORWARD AND INVERSE MODEL.

		forward model	inverse model
ESN	training	0.012	0.045
	generalization	0.019	0.049
ELM	training	0.020	0.042
	generalization	0.026	0.048

III. REFINING THE FORWARD AND INVERSE MODEL

The previous section showed that ESNs are suitable to learn the forward and inverse model of the articulatory-acoustic mapping. However, training requires a large amount of supervised articulatory-acoustic examples. In the following we use only a small subset of articulatory-acoustic data for

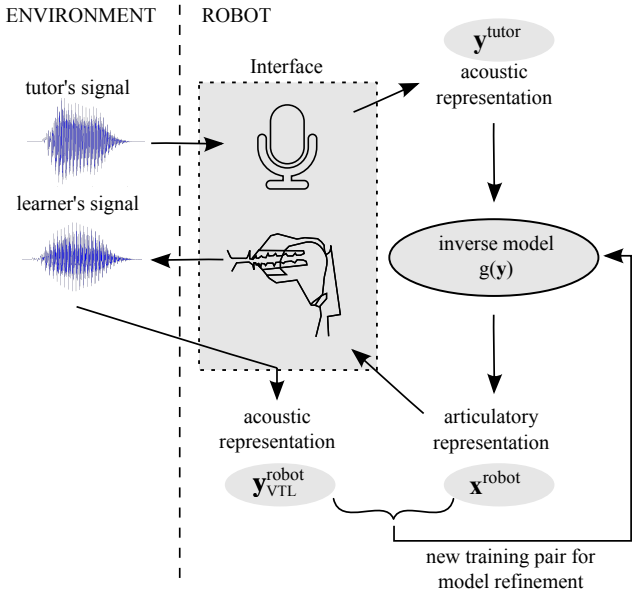

Fig. 2. The refinement process: The robot perceives a tutor's utterance, and maps the acoustics to an articulatory estimation. By producing the utterance corresponding to the estimated articulation and perceiving it, the robot receives a new training sample $(x^{robot}, y^{robot}_{VTL})$ which it uses to enhance its internal model.

supervised training. The obtained initial models are then refined by presenting new acoustic sequences *without their articulatory counterparts* to the network.

The process of refinement is depicted in Fig. 2 using the example of a robot interacting with its environment. The novel acoustic utterances are mapped to the articulatory space using the current inverse model. This mapping corresponds to an estimation process of how the perceived acoustics could be imitated. The estimated articulatory sequence is then fed into the true forward model given by the articulatory synthesizer. The generated acoustic feature trajectory y^{robot}_{VTL} together with the estimated articulatory trajectory x^{robot} then represents a new training pair which is used to update the internal forward and inverse model.

In the following, we show that the refinement process significantly enhances the inverse as well as the forward model with respect to imitation accuracy of the acoustic stimuli. Note that the learned forward model f is optional in this approach. While we assume the presence of an initial model trained on supervised data, we show that a small initial training set is sufficient. Generating initial models without requiring supervised data is subject of future work.

A. Training of the Initial Models

The initial models are trained on 8 of the 64 sequences such that every consonant and every vowel is contained once. Specifically we use the sequences [a:ba:], [a:de:], [a:gi:], [a:zo:], [a:pu:], [a:tæ:], [a:kœ:], [a:ʃy:] for training. Note that these sequences contain each consonant only in combination with a single ending vowel. The unsupervised refinement targets at improving the generalization of the models to novel contexts, where the consonants are followed by the other vowels.

The ESN parameters are chosen like in Sec. II-C. By training the initial model with a different number of variations of each of the 8 sequences we can affect its quality. In the following the number of initial training data per sequence is referred to as S. If S=1, the initial model is trained with only 1 sample of each of the 8 sequences, while S=50 refers to an initial model trained with 50×8 sequences.

B. Imitation-based Refinement

After the initial forward and inverse models have been trained, we improve the models by conducting a number of refinement iterations given by $max\_iterations$ (cf. Algorithm 1). In each iteration, 64 sequences are randomly chosen (one from each sequence class) and presented to the network. 8 of these sequences are known to the network from the initial model training. The other 56 sequences are new to the learner, as the vowels and consonants appear in novel contexts.

The learner tries to imitate the perceived acoustics by applying the current inverse model g and producing the acoustics corresponding to the articulatory estimations using the true forward model f_{VTL}. Then, error values are computed for evaluation purposes. Finally, the forward and inverse model are updated with the new training pair: the learner's articulatory estimation and the acoustic outcome.

Algorithm 1 Refinement

Require: true forward model f_{VTL}, initial forward model f, initial inverse model g
for $iteration = 0 \ldots max\_iterations$ **do**
  $\mathbf{y}^{tutor} \leftarrow$ receive new acoustic samples
  $\mathbf{x}^{robot} = g(\mathbf{y}^{tutor})$
  $\mathbf{y}^{robot}_{VTL} = f_{VTL}(\mathbf{x}^{robot})$
  $inverse\_error = MSE(\mathbf{y}^{tutor}, \mathbf{y}^{robot}_{VTL})$
  $forward\_error = MSE(\mathbf{y}^{robot}_{VTL}, f(\mathbf{x}^{robot}))$
  Update f and g with $(\mathbf{x}^{robot}, \mathbf{y}^{robot}_{VTL})$
end for

We adopt the training procedure for the output weights $\mathbf{W}^{\text{out}}$ of the Echo State Network in order to account for the sequential arrival of new data in the refinement phase. We apply the online sequential learning introduced for Extreme Learning Machines in [20] to the ESN, which proceeds similar to recursive least squares [21].

C. Evaluation of the Forward and Inverse Model Refinement

Errors of the forward and inverse model are measured in the acoustic domain as illustrated in Fig. 3. The inverse model error is calculated between the initial acoustic sequence and the signal reproduced from the estimated articulatory sequence via *VocalTractLab*, i.e. it is the difference between the tutor's signal and the learner's imitation.

To evaluate the forward model, we take the estimated articulatory sequence $\mathbf{x}^{robot}$ as a basis and compare the outcome of the learner's internal forward model $\mathbf{y}^{robot}$ to the acoustics generated by the true forward model $\mathbf{y}^{robot}_{VTL} = f_{VTL}(\mathbf{x}^{robot})$. The forward model error therefore expresses how well the learner can predict the acoustic outcomes of its vocal tract.

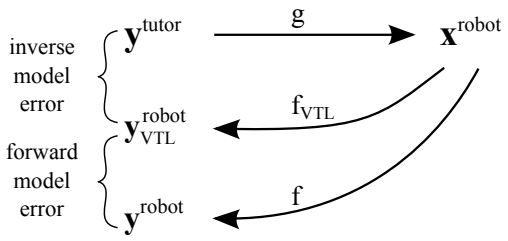

Fig. 3. Measurement of the forward and the inverse model errors in the refinement process.

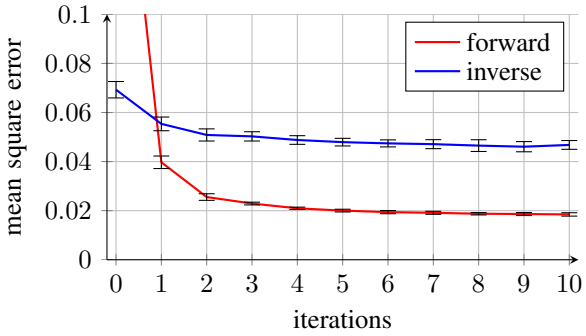

Fig. 4. Mean and standard deviation of the forward and inverse model errors of the initial model trained with 3×8 sequences and during refinement.

D. Refinement Results

Tab. II and Tab. III show the performances of the initial models and the refined models for the forward and the inverse mapping direction, respectively. For each value of S the errors were averaged over 5 repetitions of the experiment with $max\_iterations = 10$ iterations of refinement.

The performance of the initial model depends highly on the number of initial training data: Models trained with a larger amount of data produce lower errors. The refinement process reduces the error of the forward and inverse model in all cases significantly. While the error decreases quickly in the first iterations, it converges to a minimum that is comparable to the earlier generalization results presented in Tab. I. The standard deviations of the errors in the last iteration are very low (on average 0.0008 for the forward model and 0.0025 for the inverse model).

After 10 iterations of refinement, the lowest errors of the forward and inverse model can be found for S=3. But even for minimal initial models with S=1, the errors decrease to a low level in both models. For models with a higher number of initial training data it can be observed that the errors after 10 iterations are slightly larger than for starting with a smaller S. This is due to the fact that the influence of newly arriving data is weaker in the beginning and gets balanced with respect to the initial training data during prolonged iteration. After 50 iterations, the S=50 model as well reaches an error of 0.019 for the forward and 0.047 for the inverse mapping which is comparable to the errors of the S=3 models after 10 iterations. Thus, refinement succeeds independent of the number of initial training data.

In Fig. 4, the forward and the inverse model errors together with their standard deviation are plotted over the refinement process for an initial model with S=3. It can be noticed that especially the error of the forward model is very high in the initial model and decreases by approximately 90% during the refinement. The reason for the high initial error is that the forward model is applied on the estimated articulatory sequences which may be initially very different from the known articulatory trajectories. After the first iteration, the prediction error of the forward model decreases drastically as the network now contains acoustic correspondences for such estimated articulatory sequences.

All in all, the results demonstrate that the refinement process decreases the forward and inverse model errors significantly. An initial model trained with only 3×8 articulatory-acoustic sequences is sufficient for the proposed refinement strategy to reach generalization errors similar to the cross validation results on the complete data set with 64 sequence classes, while requiring only acoustic data.

Tab. II
FORWARD MODEL ERRORS FOR 0 . . . 10 ITERATIONS OF REFINEMENT AND DIFFERENT SIZES S OF THE INITIAL TRAINING SET

	Forward error after . . . iterations					
S	0	1	2	5	10	Error decrease
1	0.846	0.055	0.035	0.025	0.022	97.4%
3	0.208	0.040	0.026	0.020	0.018	91.1%
5	0.164	0.040	0.027	0.021	0.019	88.4%
10	0.160	0.039	0.029	0.022	0.020	87.8%
50	0.165	0.044	0.036	0.029	0.025	85.1%

Tab. III
INVERSE MODEL ERRORS FOR 0 . . . 10 ITERATIONS OF REFINEMENT AND DIFFERENT SIZES S OF THE INITIAL TRAINING SET

	Inverse error after . . . iterations					
S	0	1	2	5	10	Error decrease
1	0.098	0.071	0.065	0.059	0.057	41.4%
3	0.069	0.055	0.051	0.048	0.047	32.5%
5	0.066	0.055	0.053	0.051	0.049	25.5%
10	0.063	0.053	0.052	0.049	0.048	23.8%
50	0.062	0.056	0.054	0.052	0.050	18.9%

E. Perceptual Evaluation

In addition to the error-based analysis, the authors also qualitatively evaluated the results by listening to the acoustics corresponding to the estimated articulatory trajectories of the inverse model (as reproduced by *VocalTractLab*). They listened to each of the 64 syllables before refinement and after 10 iterations of refinement in comparison. Then they rated which sample is better comprehensible or if the comprehensibility does not change. An initial model trained with 3×8 sequences was chosen.

This perceptual evaluation revealed that approximately 40% of the sequences can be better recognized after the refinement, while 12% become less comprehensive. For the others no significant difference can be heard before and after the refinement. The most common improvement is that the acoustic sequences estimated by the initial inverse model sometimes contain click noises, but sound more smooth and natural after the refinement.

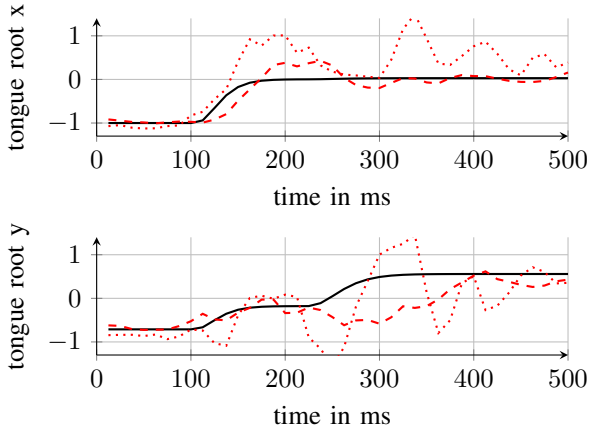

Fig. 5. Horizontal (top) and vertical (bottom) tongue root position of the utterance [aːgæː]. The black line is the ground truth, the red lines are the estimations of the initial (dotted line) and the refined model (dashed line).

As an example, Fig. 5 shows the horizontal and vertical tongue root position parameters of the sequence [aːgæː] for an initial model with S=1. The position of the tongue root is important for the correct reproduction of the consonant [g]. The black lines show the articulatory trajectories that were used to generate the acoustic target. The red dotted lines show the articulatory trajectories as estimated by the initial model, and the red dashed lines are the articulatory trajectories estimated by the refined inverse model.

It can be observed that the reproduction of the initial model is very rough and noisy. Such noisy outcomes can be especially observed in case of few initial training data. After model refinement, the articulatory trajectories are smoother and more accurate. Although the estimation is still not perfect, this improvement leads from an incomprehensible utterance to a clear [aːgæː] and thus a successful acoustic imitation using the refined inverse model.

Not all sequences are improved during the refinement. Especially those sequences used for initial training can be better approximated by the specialized initial models than by the refined models, which is an expectable result. 48% of the sequences do not show a perceivable improvement at all. The problem is two-fold: Firstly, in the learning process errors are weighted equally in each articulatory parameter, whereas perceptual features in fact change in a highly non-linear manner with respect to articulatory parameter changes. This can be addressed by utilizing respective error metrics, which, however, are not easy to define. Secondly, the unsupervised refinement presented in this paper does not actively explore its actuation space in order to achieve better imitation results. To complement the refinement with an active exploration mechanism is subject of future work.

IV. Conclusion

We showed the efficient learning of forward and inverse models for speech production and imitation using a recurrent neural network model. The considered data contain coarticulations of a large set of vowels and consonants. We demonstrated that initial models trained on small subsets of articulatory-acoustic data can be improved significantly by imitation-based refinement. This unsupervised process requires only acoustic data and can be developmentally interpreted as imitative learning in a tutoring situation.

Acknowledgment

This research was funded by the DFG, Cluster of Excellence 277 "Cognitive Interaction Technology" and is related to the European Project CODEFROR (PIRSES-2013-612555).

References

[1] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3, pp. 303–319, 2002.

[2] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping." in *Interspeech*, 2006.

[3] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[4] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 118-1, pp. 444–460, 2005.

[5] A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research." *Phonus.*, vol. 5, 2000.

[6] G. Westerman and E. R. Miranda, "Modelling the development of mirror neurons for auditory-motor integration," *Journal of New Music Research*, vol. 31, no. 4, pp. 367–375, 2002.

[7] Y. Yoshikawa, J. Koga, M. Asada, and K. Hosoda, "Primary vowel imitation between agents with different articulation parameters by parrot-like teaching," in *IEEE/RSJ Intern. Conf. on Intelligent Robots and Systems*, 2003, pp. 149–154.

[8] Y. Sasamoto, N. Nishijima, and M. Asada, "Towards understanding the origin of infant directed speech: A vocal robot with infant-like articulation," in *IEEE Joint Intern. Conf. on Development and Learning and Epigenetic Robotics*, 2013.

[9] A. S. Warlaumont, G. Westermann, E. H. Buder, and D. K. Oller, "Prespeech motor learning in a neural network using reinforcement," *Neural Networks*, vol. 38, pp. 64–75, 2013.

[10] C. Moulin-Frier and P.-Y. Oudeyer, "Exploration strategies in developmental robotics: a unified probabilistic framework," in *IEEE 3rd Joint Intern. Conf. on Development and Learning and Epigenetic Robotics*, 2013, pp. 1–6.

[11] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and language*, vol. 96, no. 3, pp. 280–301, 2006.

[12] B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, vol. 51, no. 9, pp. 793–809, 2009.

[13] C. Moulin-Frier, S. M. Nguyen, and P.-Y. Oudeyer, "Self-organization of early vocal development in infants and machines: the role of intrinsic motivation," *Frontiers in psychology*, vol. 4, 2013.

[14] M. Jordan and D. Rumelhart, "Forward models: Supervised learning with a distal teacher," *Cognitive Science*, vol. 16-3, pp. 307–354, 1992.

[15] N. Ueffing, G. Haffari, and A. Sarkar, "Transductive learning for statistical machine translation," 2007.

[16] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 216–229, 2010.

[17] P. Birkholz, "VocalTractLab – Towards high-quality articulatory speech synthesis," http://www.vocaltractlab.de/.

[18] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks," German National Research Center for Information Technology, Tech. Rep. 148, 2001.

[19] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *IEEE Intern. Joint Conf. on Neural Networks*, vol. 2, 2004, pp. 985–990.

[20] N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1411–1423, 2006.

[21] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 1991.