Modeling Speech Imitation

Jonas Hörnstein[1], Lisa Gustavsson[2], José Santos-Victor[1], Fransisco Lacerda[2],

[1] Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa, Portugal
[2] Department of Linguistics, Stockholms University, Stockholm, Sweden

1 Introduction

The concept of imitation is often pointed out as one of the cornerstones in infants' early language acquisition. Still there are few studies concerning vocal adult-child imitations reported in the literature and results from those are often inconsistent. One reason for the inconsistencies is the lack of a stringent model for what should be classified as imitations. This is not only a problem when trying to learn something about adult-child interactions, but also when trying to make robots that can learn to interact naturally with humans. In order to interact vocally a robot is typically equipped with artificial models of the ear and the vocal tract connected by an artificial neural network. This model is inspired by the motor theory of speech perception [1] and the more recent discovery of mirror neurons [2]. While the robot can use babbling to create an initial map between the acoustic signal and the corresponding vocal tract positions, it needs to overcome inter-speaker differences and to acquire key positions of the vocal tract to be able to communicate with humans or other robots. Imitation games are thereforee used to train the networks [3, 4, 5]. As we have shown in our previous work [6] these imitation games should preferably go both ways. Having the robot imitating the caregiver is useful for directing the robot towards key-points, while having the caregiver imitating the robot is more important for learning the map and overcome inter-speaker differences. As we will show in this work, both types of imitations can also be found in adult-child interactions. However, while robots usually follow very strict imitation games with predefined turn-taking behaviors, adult-child interactions tend to be much more complex. For the robot to be able to learn its maps under such natural conditions it has to be able to separate imitations from non-imitations. The question we want to answer in this work is therefore the following. How can the robot decide when a pair of utterances should be considered as vocal imitations of each other?

To answer this question we first make an imitation judgement experiment where a panel of listeners get to classify utterances recorded during natural adult-infant

interactions. Second we derive a number of tonotopic speech representations and use those to create a classifier that separates the utterances into imitations and non-imitations. Third we test the performance of the classifier in a human-robot interaction experiment.

2. Imitation judgement experiments

Data for the judgement experiment was recorded during 15 half-hour sessions. Seven Swedish infants, with ages ranging from 185 to 628 days, participated in one, two or three sessions each. A lot of care was taken to allow for natural interactions during the experiments. The recordings were made in a comfortable home-like environment, in a recording studio at the Phonetics Laboratory, Stockholm University. The infant and the adult were free to move around during the recordings and they were also provided with a number of toys. In total, these recordings generated an adult-infant interaction speech data base consisting of 4100 speech samples.

A computer program created in LabView was used to select utterances from the database and present those for the user. The program randomly draw an utterance from the pool of adult utterances and then randomly selected the utterance that the infant produced within five seconds before or after the adult's utterance. For the imitation judgement 20 subjects were each presented with 150 pairs of utterance (50 from each age group). Of these the subjects evaluated 22% as imitations, 19 % as uncertain, and 59 % as non-imitations.

There were no significant difference in the number of perceived imitations between the three age groups. However, there was an obvious difference in the way subjects classified a pair of utterances based on the perceived age of the infants. The older the infant are the higher are the demands on matching parameters. This is nothing new and can be illustrated with the familiar example of [baba] that happily is rewarded as an imitation of both [mama] and [papa] when the infant is very young, but it is still worth to mention as it have implications on the way we chose to build our classifier. Further analysis of the results showed that in 52% of the cases that a pair of utterances were judged as an imitation it was the adult imitating the infant.

3. Imitation classifier

To compare the acoustic signals of the infant and the adult in each utterance-pair the robot needs a set of tonotopic sound representations or sound features that it can use for the classification. Because of anatomical and physiological differences between adults and infants there will obviously never occur a perfect match in the acoustic realizations of the utterance-pairs. We therefore want to avoid a direct comparison of representations such as the pitch and MFCC. Instead we created a number of features based on the temporal envelope and temporal differences of the spectral features. These were:

1. Number of syllables
2. Length of the last syllable
3. Length of the second last syllable
4. Difference in length between the two last syllables
5. Difference in pitch for the two last syllables
6. Difference in the first MFCC of the two last syllables

For the number of syllables we simply classified every utterance-pair where the number of syllables of the infant and the adult utterance did not match as a non-imitation. Such a straightforward classification could not be done for the other features. Even for a true imitation we cannot expect the difference between the feature values of the two utterances to be zero. We used a gamma-distribution to model the differences for each feature in the case of an imitation and a non-imitation. The parameters of the gamma-distribution were then estimated from the normalized distance between the feature values for each pair of utterances. This was done separately for utterances judged as imitations and non-imitations, as well as for each age group.

In order to mimic the way human listeners seemed to incrementally demand a more detailed match between the utterances as infants got older, we took an hierarchical approach by adding an extra classifier for each age group. For the youngest infants the length of the last syllable showed to be the most efficient feature for separating imitations from non-imitations. We chose the crossing between the two distributions in Figure 1 as a separation point for when an utterance should be classified as an imitation. Doing so 74% of the utterances classified as imitations had also been classified as imitations by the panel subject, while 26% were false positives. Using the same feature for the data in the second age group, we only get around 50% true classifications. However, when adding a second classifier based on the fourth feature,

i.e. difference in length between the two last syllables, the combined classifier gave around 80% true positives for the second age group and around 65% for the third age group. Finally we added a third classifier based on the difference in pitch which was able to completely eliminate the false positives for the third test group.

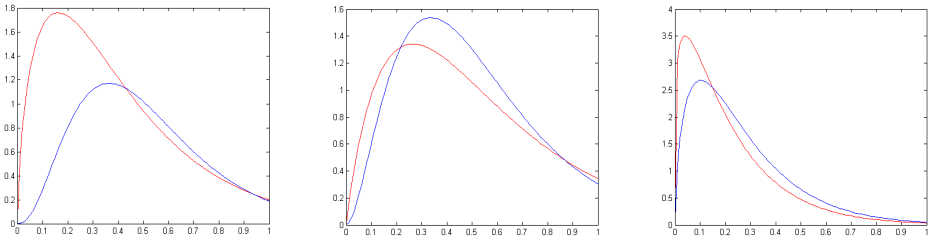

Figure 1. Expected feature values for imitations (red) and non-imitations (blue). The plot show the features used for the first, second, and third classifier respectively.

4. Human-robot interaction experiment

Finally the imitation classifier was tested during a simple human-robot interaction game, where the robot randomly selected a number of points in the vocal tract model and created trajectories between those. The utterance was then synthesized and played for the caregiver who tried to imitate the sound. When presenting the utterance from the robot and the caregiver to the classifier 66% were classified as imitations by the first classifier, 39% by the second, and only 3% by the third classifier.

We also tested how important it is for the robot to be able to detect false imitations when trying to learn the map between human speech sounds and its own vocal tract by doing a simulated imitation game. For this experiment we used a database of Portuguese vowel sounds from 14 different speakers and the setup described in [6]. Data from seven of the speakers were used for training the network and data from the other seven were used for testing. The robot were presented by a mix of correct vowel imitations and false positives. First we tested how well the robot could learn the map without checking if it is a true imitation by using only 22% correct data. Then we simulated the use of the hierarchical classifier and first trained with 72% correct data, followed by 80%, and finally 100% correct data. Figure 2 shows how the error in the audio-motor map is reduced during the different training steps. It can be seen that the error in the map reduces quickly in the beginning of the training cycle, even in the

case of many false imitations. However, without the use of the classifier the error remains at a relatively high level, even after several iterations. When using the hierarchical classifier the error of the map is gradually reduced with each step. To get a better understanding of how well the different maps work in practice we let the robot try to reproduce the sounds in the test set using each map. Even if the map trained without the classifier was able to produce some vowel sounds it was obvious that the maps trained with the classifiers were gradually producing better results. This became even more clear when trying to classify the nine vowels uttered by the people in the test set based on the mapped vocal tract positions. When using the map trained without the classifier the recognition rate was as low as 27%. With the hierarchical classifier the recognition rate became .48% using one classifier, 56% using two classifiers, and finally 58% using all three classifiers

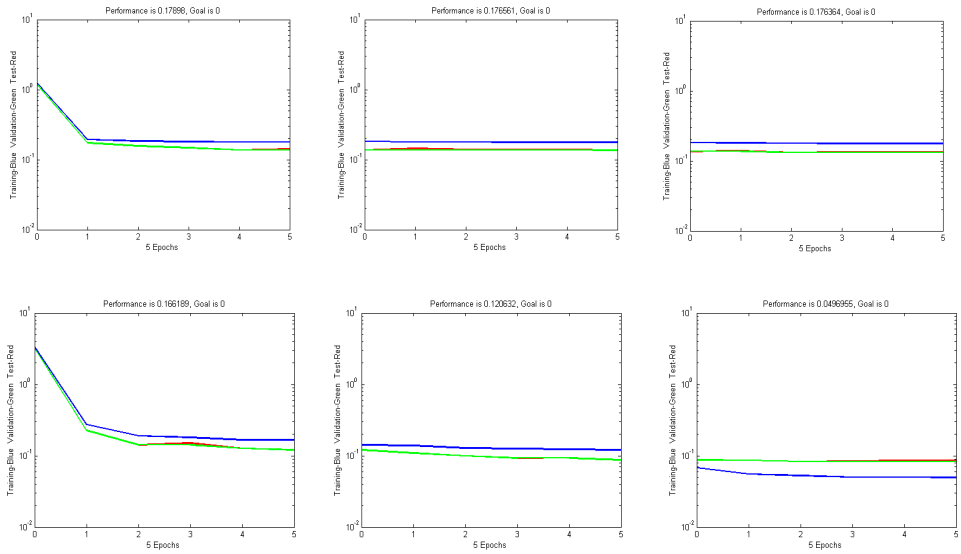

Figure 2. Learning the sound motor map during a simulated interaction game. Above without a classifier during all three development stages, below using the hierarchical classifier.

5. Discussion

In this work we have identified the need for being able to separate between imitations and non-imitations if a robot should be able to acquire a language through natural interactions with a caregiver rather than through highly structured imitation games as

is usually the case today. We suggest the use of features based on the prosody of each utterance in an utterance-pair rather than directly comparing spectral features in order to overcome anatomical and physiological differences between the speakers. We also suggest a developmental approach where the classification of an utterance-pair as an imitation or non-imitation is based on a single feature in the early developmental stage and then gradually gets based on more information in the later stages. This seems to coincide with how human listeners classify the same adult-child utterance-pair as either an imitation or a non-imitation based on the developmental stage of the child. To realize this we use an hierarchical classifier.

Initial experiments show that by using few features in the early developmental stage the child or robot is quickly able to create an initial map which can be fine tuned during the later developmental stages by adding more features.

References

1. Liberman, A. and Mattingly, I., “The motor theory of speech perception revisited”. Cognition, 21:1–36. (1985)
2. Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. “Speech listening specifically modulates the excitability of tongue muscles: a tms study”, European Journal of Neuroscience, Vol 15:399–402. (2002)
3. de Boer, B., The origins of vowel systems, Oxford Linguistics, Oxford University Press. (2000)
4. Oudeyer, P. Y., “The self-organization of speech sounds”, Journal of Theoretical Biology, (2005)
5. Guenther, F. H., Ghosh, S. S., and Tourville, J. A., "Neural modeling and imaging of the cortical interactions underlying syllable production", Brain and Language, 96 (3), pp. 280-301
6. Hörnstein, J. and Santos-Victor, J. (2007). A unified approach to speech production and recognition based on articulatory motor representations. In IROS07, pages 3442–3447. (2007)