Using Imitation to learn Infant-Adult Acoustic Mappings

G. Ananthakrishnan, Giampiero Salvi

Center for speech Technology,
KTH (Royal Institute of Technology), Stockholm, Sweden
agopal@kth.se, giampi@kth.se

Abstract

This paper discusses a model which conceptually demonstrates how infants could learn the normalization between infant-adult acoustics. The model proposes that the mapping can be inferred from the topological correspondences between the adult and infant acoustic spaces, that are clustered separately in an unsupervised manner. The model requires feedback from the adult in order to select the right topology for clustering, which is a crucial aspect of the model. The feedback is in terms of an overall rating of the imitation effort by the infant, rather than a frame-by-frame correspondence. Using synthetic, but continuous speech data, we demonstrate that clusters, which have a good topological correspondence, are perceived to be similar by a phonetically trained listener.

Index Terms: infant speech acquisition, unsupervised learning, self organizing maps

1. Introduction

An infant learning to communicate with speech has been a source of intrigue and interest, both from the point of view of psychology and medicine, and from the point of view of speech research. Understanding this phenomenon is especially difficult, given that infants do not remember the process once they grow up and there may be limited means of communicating with infants while the process of learning takes place.

There are several challenges an infant faces when trying to acquire the ability to speak. The first and the most important challenge is learning the sensorimotor mapping between the acoustics and the articulatory configurations [1, 2, 3]. Most of the theories in the above studies make use of the phenomenon of babbling to explain this process. Secondly, an infant learning to speak needs to learn how to categorize the different sounds that he/she produces or hears from the adults [4, 5].

An infant also needs to understand how to correlate the different sounds produced during babbling to the sounds present in adult speech which is the main focus of our study. There have been several proposals about which acoustic correlates are invariant between infant and adult speech (e.g., [6]). These measures were evaluated by [7] and [8] and were shown to be useful in acquiring the ability to imitate the sounds produced by adults. However, are these invariant acoustic features that normalize adult speech, intrinsically and instinctively known to children, or do they learnt it? While most studies have assumed this to be innate to children, Plummer *et.al* [9] have proposed a method which models the speaker normalization acquisition for several languages. The correspondence between the adult and child renderings of a few typical vowels was assumed to be available for learning the mapping between the two vowel systems. It was shown that given a few corresponding instances, using a semi-supervised method based on topological alignment of the manifold, the mapping could be learnt for the entire vowel system. However, a model which requires the knowledge of correspondences between certain exemplars in the adult and infant speech, is not a realistic scenario for infants. Besides, the study was also restricted to stationary vowels produced both by the adults and the infants.

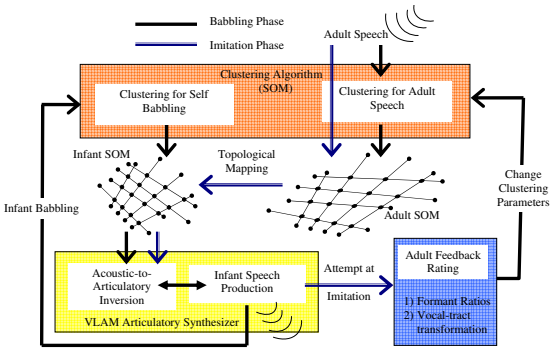

Figure 1: The block diagram of our model for acquiring speaker normalization. It shows the paths for both the babbling phase and the imitation phase.

This paper attempts to continue on the lines of [9] in order to investigate how children learn the appropriate features that help in normalizing between adult and their own voices, in spite of the seemingly large differences in the acoustic spaces. We propose a method to learn the mapping without the necessity of knowing specific corresponding exemplars. The important assumption of maintaining the geometric relationships in the adult and infant acoustic spaces is retained in our study. Instead of example-by-example supervision, the infant only receives an overall feedback from the adult about the overall imitative babbling endeavor made by the child. This feedback shapes the learning process, even though ratings or correspondences of individual frames of speech is not available. This, according to us, is a more realistic scenario in the infant speech acquisition paradigm.

2. Experimental Paradigm and Tools

Our model is restricted to studying the acquisition of speaker normalization skills. We do not focus on the other aspects of infant speech acquisition and development. For this reason, we make certain assumptions and simplifications. The block-diagram of our model is shown in Figure 1. The various tools that the model uses, are explained in the following sub-sections.

2.1. The Articulatory Synthesizer

We use a Variable Linear Articulatory Model (VLAM) [10] with a reduced length of the oral cavity as against that of the phar-

ynx, obtained from [11]. The present study assumes vocal-tract length and dimensions corresponding to that of a 10 month old infant. The synthesizer has 7 articulatory parameters, which are controlled from values ranging from -5 to 5 Standard Deviations (SD) from the mean.

2.2. Topological Clustering

The motivation behind performing a topological clustering is the assumption that even though the transformation from the infant acoustic space to the adult acoustic space may not be easy to estimate, with the transformations being locally non-linear, the topology of the two spaces would still remain similar. Self Organizing Maps (SOM) [12] perform an unsupervised clustering while maintaining the topological constraints on the original data. By using SOM with the same set of parameters to cluster both the adult and infant acoustic spaces separately, we hope to be able to make a topological mapping between the two spaces. Each cluster of the adult acoustic space should correspond to a cluster in the infant acoustic space, which lies in the same topological location. The problem, however, is to find the correct parameters of the topology for the clustering. This is described in Section 2.5.

2.3. Simulation of Babbling

Babbling has often been considered very important for learning the mappings between acoustics and articulation. We consider this phase important even for learning speaker normalization. We generate infant babbling by manipulating the articulatory parameters. We pick articulatory targets for each articulatory parameter at random, to be achieved at certain randomly selected timings. The articulatory path is then calculated, by applying a minimum jerk trajectory between the two articulatory targets. In reality, there are some patterns that babbling infants follow, and the targets are not truly random. The increasingly complex patterns of babbling are motivated by the language the infant hears from the adults. This is, however, beyond the scope of the current paper.

The babbling sounds that are created by our synthesizer include all vowel like, as well as other sonorant-consonant like sounds. Voiced stop consonants are also produced, but do not form a large percentage of the produced sounds. Fricatives and aspirated stop consonants are not produced in this study due to the limitations of the synthesizer we adopted.

We use synthesized adult data from a female adult vocal-tract model for training our models, instead of real recorded data, for two reasons. The first is so as not to rely on formant detecting algorithms which perform rather poorly on infant voices. The second reason is to simplify the simulation of adult assessment and feedback on the quality of the imitation by the infant. Knowing the vocal-tract information of the target utterance helps in making a transform from the infant voice to the adult voice easier. A detailed explanation can be found in Section 2.6. The adult utterances we use are produced by the same babbling procedure described for the infant, making it language independent. This simplification, although reasonable for this study, is unrealistic in the sense that the true adult speech will be influenced by the languages they speak.

2.4. Acoustic Features

We use two sets of acoustic features. The first one is 13 Mel Frequency Cepstral coefficients (MFCC). These are not normalized for different vocal-tract lengths. The infant needs to learn a method to transform the MFCCs corresponding to its own voice to the MFCCs produced by adults. Since the data produced contains both sonorants and stop consonants, formant frequencies are unavailable at all frames of the data. Instead, MFCCs allow modeling of both vowels as well as consonants. The SOM, which simulates the topological clustering, uses these MFCC features to cluster both the acoustic spaces.

The second acoustic feature that is used is the first three Formant Frequency ratios, proposed by [6], as a speaker invariant acoustic measure. The three acoustic features are

$$\left[\log\frac{F1}{SR}, \quad \log\frac{F2}{F1}, \quad \log\frac{F3}{F2}\right] \tag{1}$$

where $F1$, $F2$ and $F3$ are the first three Formants, while SR is calculated from the geometric mean of the fundamental frequency $mF0$ as $SR = 168\,(mF0/168)^{1/3}$. It has been shown [8] that using formant differences (formant ratios in the log domain) is one of the better acoustic measures in order to simulate imitative babbling by infants. We propose that this measure may be unknown to the infant, or rather is the measure that the infant must learn. This measure is used by us only in calculating the adult feedback (assuming that this measure is known to the adults), and not for modeling the infant learning process.

2.5. The process of learning speaker normalization

As an infant is exposed to different sounds in its environment, produced both by itself and the adults around, the infant tries to cluster these sounds into groups. However, instead of classifying all the sounds into the same groups, we propose a model where the infant performs a clustering of sounds produced during babbling, separately from the clustering of sounds produced by the adults. The crux of our model is this clustering, which needs to be topological in nature, and the same number of clusters need to be estimated for both infant and adult voices. One of the main issues is that the infant does not know the correct number of clusters. If the infant performs clustering with too few clusters, then, even though the cluster topological correspondence would be correct, the imitative babbling will produce too few sound types, increasing the imitation error. On the other hand, if there are too many clusters, the topological invariance assumption will no longer hold, since the true mapping is a non-linear one. This will result in topology correspondence errors. Another problem would be, if the infant uses an incorrect topological structure in terms of the the number of dimensions and the shape of the topology in order to cluster the two acoustic spaces.

This is problem is solved in our model during the imitation phase. The infant uses adult feedback on its imitation performance (detailed in the following section) to select a suitable size and structure for the cluster topology. A topological structure which enables a better imitation, is more suitable for learning the mapping. During imitation, the infant first classifies the target adult utterance into the adult clusters, finds the topological mapping between the adult clusters and infant clusters, performs acoustic-to-articulatory inversion (henceforth, inversion) to find the articulatory parameters and then proceeds to imitate the adult. The process of inversion is not simulated perfectly by our study. Instead, the mean values of the vocal-tract area functions parameters from all the training data, classified under each cluster, is used to produce the imitative speech. A minimum jerk smoothing is performed on the articulatory parameters before producing the imitation.

2.6. Simulating Adult Feedback

Adult feedback to the infant is extremely limited and non-specific, for the reason that speech based communication is not possible with the infant. For this reason, models which assume

that one-one correspondence between adult and child productions is available to the infant, are highly unrealistic. What the infant will receive is an overall rating on its attempt at imitating the adults.

In trying to simulate the adult feedback, we make use of a global measure on a number of imitative productions in terms of the Root Mean Square Error (RMSE) of production. Comparing the MFCCs of infants productions directly with the target adult utterances would not be useful due to the mismatch between adult and infant voices, giving a poor rating to the imitation effort. We assume that the adults know what transform to perform on the infant speech in order to rate it. We simulate adult feedback using the following methods

1. We used features suggested by [6], applicable only to vowels. These features provide an error metric that is speaker invariant and thus provide a good means for feedback to the infant. In this study, frame-by-frame feedback or even utterance-by-utterance feedback is not available to the infant. The overall feedback is available over several imitation utterances. So regions where formants are undefined are not taken into account.

2. Since we know the exact dimensions of the adult vocal-tract that produced the utterance, as well as the vocal-tract dimensions of the infant, we can transform the infant vocal-tract configurations to that of the adult. The speech produced by the transformed adult vocal-tract, generated from vocal-tract area functions of the infant while imitating, is then used for direct comparison with the adult speech. The RMSE is now calculated on the MFCCs, which include both vowels and consonants. We assume that this transformation is comparable to the transformation the adults make when they perceive infant speech. This rating method is valid only if one knows the exact vocal-tract configurations of the infant and the adult. For this reason, we restricted ourselves to using synthesized adult speech material for the experiments, instead of using real adult speech.

3. Experiments and Results

We generated 220 utterances each with the adult and infant voices. These consisted of around 50000 frames of speech for each voice. The frame rate used was 500 ms and sampling frequency was 8000 Hz . Each utterance contained what would correspond to 1 to 5 syllables. The fundamental frequency and the loudness of the different utterances were varied too, for both the sets of data. We then produced 22 utterances of adult speech, which are used for obtaining adult feedback. The infant is simulated to perform independent SOM clustering of the two data sets, with the same topological parameters for the two acoustic spaces. Clustering was performed with the number of clusters ranging from 1 to 60, and the topology dimensions from 1 to 3. Different shapes for the topology, including 'sheet','cylinder' and 'toroid' were also applied. The best topology was selected by minimizing the RMSE using the two error metrics described in Section 2.6.

The experiments suggested that, the best adult feedback using the features in Equation 1 were obtained using 15 clusters, in a one dimensional topology. The shape of a single dimensional topology did not matter. Using the comparison of the vocal-tract transformed imitations, the best adult feedback was obtained for 18 clusters, again in a single dimension. Figure 2 shows the performance of the topological mapping and hence the imitation with respect to the different number of clusters. Figure 3 shows the mapping in the formant frequency ratio do-

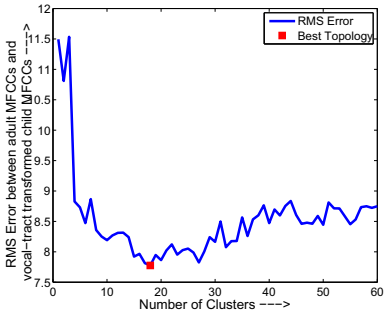

Figure 2: Illustration of the Adult feedback in terms of RMSE over 22 utterances between the original adult speech and the infant imitation produces from the transformed adult vocal-tract.

main between topologically corresponding clusters. One can see that even though the topology is more or less maintained, the error between the adult clusters and the infant clusters can be quite large. It can be observed that some topologically corresponding clusters may not find a similar correspondence in the formant ratio space, such as clusters 1, 2, and 7.

In order to verify how well the correspondences were made, the prototype sounds of the clusters were classified into the closest International Phonetic Association (IPA) symbols by a phonetically trained listener. The listener was also asked to give a rating to how well the topologically corresponding clusters corresponded perceptually. Perfect perceptual correspondence was awarded ten points and complete lack of similarity was awarded one point. Table 1 indicates classification and ratings given to the different clusters in the adult and infant spaces. It is clear that even though consonants were capable of being produced by the synthesizer, most of the clusters corresponded to vowel like sounds, probably due to their longer durations. The minimum rating of 3 and a maximum of 10 was received with a mean rating of 7.6 for the 15 clusters. A similar rating was performed for the best configuration obtained using the transformed vocal-tract lengths and the minimum rating was 4 and the maximum rating obtained was 10, with a mean rating of 7.9 for the 18 clusters.

Table 1: *Ratings by the phonetically trained listener on the correspondence between the infant and adult topologically mapped clusters for the best topological configuration.*

Clust. No.	Closest IPA sym. (adult)	Closest IPA sym. (infant)	Rating (best=10)	RMSE (log freq.)
1	œ	ɜ	5	0.2
2	ɔ	e	5	0.27
3	i	i	8	0.37
4	j	j	9	0.42
5	y	y	7	0.36
6	ø	ø	9	0.08
7	ɵ	ɛ	3	0.21
8	ʊ	ɵ	7	0.25
9	o	ɔ	8	0.32
10	ɐ	æ	8	0.15
11	ɛ	ɛ	10	0.09
12	a	a	9	0.15
13	æ	æ	9	0.38
14	ɑ	ɑ	9	0.29
15	ɑ	œ	8	0.27

The perceptual ratings by the listener shows similarities to the RMSE between the adult and infant clusters both in the log formant frequency space as well as the vocal-tract transformed space. In general, larger distances between the clus-

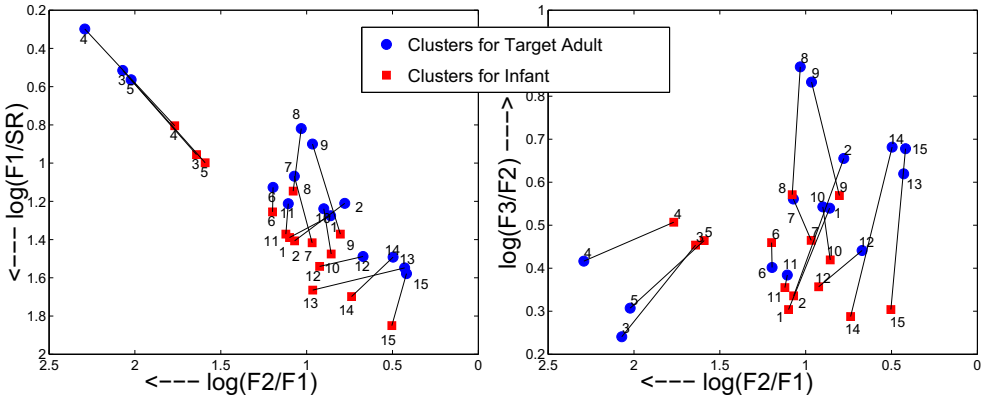

Figure 3: Figure showing the topologically corresponding adult-infant clusters in the formant ratio space. One can see that although most of the clusters are topologically compatible to each other, some clusters are spaced quite wide apart, resulting in larger errors.

ters had a lower perceptual rating, though not the case for all clusters (*cf.* Table 1). This shows that the features we used may not correspond perfectly to perception of speaker normalization, although more information is needed to make definite conclusions.

4. Conclusion and Future Work

This paper intends to model the acquisition of infant-adult speaker normalization by an infant, by performing a clustering in infant and adult acoustic spaces, separately, and then using the topological correspondence between them. SOM, an unsupervised clustering method which preserves the topology of the data was used for this purpose. Adult feedback in terms of RMSE either in the formant ratio space or between the MFCCs of the original adult speech and the vocal-tract transformed infant imitation was used to select the best topology for the clustering. Perceptual ratings by a phonetically trained listener showed a reasonable correspondence between the clusters formed in the infant space and adult spaces.

This study although by no means claims to solve the problem of learning speaker normalization, provides a novel model and insight into the problem. The method can be applied to any class of phonemes. There are several improvements one could make to the model. The utilization of adult feedback is limited to selecting a suitable topology, and not in the clustering itself. Although the simulations are based on continuously varying speech data, the clustering performed in this study does not take into account the dynamic aspects of speech and each frame is considered independent of the previous and succeeding ones. A clustering taking these aspects into account, would provide better normalization results. The inversion method used was extremely simple, which may cause errors in the simulation of imitation. Using a more sophisticated inversion method would probably improve the quality of imitation.

The main drawback for the current study is the use of synthesized adult speech. Since the only reason why real speech is not used is to be able to provide feedback, the model should be valid for real speech in principle. Our immediate future work is to generate the clusters from real speech, but perform the imitation evaluation on synthesized speech. The simulations in this study are restricted to a 10 month old infant. Tying the number of clusters, imitation performance as well as the adult feedback to the growth in age of the infant would be interesting future work.

As an offshoot to this model, one may predict that the best correspondences between adult and infant speech will occur when the infant learns to cluster both the spaces in the same manner as the adult, i.e., into the phonemes of the language. Demonstrating this would require performing this experiment on adult speech corresponding to a specific language, something we envisage doing in the future.

5. Acknowledgements

This research was partly funded by the Swedish Research Council (Grant No. 2009-4599).

6. References

[1] Guenther, F., "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production," Psychological Review, 102(3):594–620, 1995.

[2] Markey, K., *The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development*, Ph.D. thesis, University of Colorado, 1994.

[3] Bailly, G., "Learning to speak. Sensori-motor control of speech movements," Speech Communication, 22(2-3):251–267, 1997.

[4] Kuhl, P., "Innate predispositions and the effects of experience in speech perception: The native language magnet theory," Developmental neurocognition: Speech and face processing in the first year of life, 259–274, 1993.

[5] Werker, J. and Curtin, S., "PRIMIR: A developmental framework of infant speech processing," Language Learning and Development, 1(2):197–234, 2005.

[6] Miller, J., "Auditory-perceptual interpretation of the vowel," J. of Acous. Soc. Am., 85(5):2114–2134, 1989.

[7] Ménard, L., Schwartz, J. L., Boë, L. J., Kandel, S., and Vallée, N., "Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood," J. of Acous. Soc. Am., 111:1892, 2002.

[8] Heintz, I., Beckman, M., Fosler-Lussier, E., and Ménard, L., "Evaluating parameters for mapping adult vowels to imitative babbling," in Proc. Interspeech, 688–691, 2009.

[9] Plummer, A., Beckman, M., Belkin, M., Fosler-Lussier, E., and Munson, B., "Learning speaker normalization using semisupervised manifold alignment," in Proc. Interspeech, 2918–2921, 2010.

[10] Maeda, S., "An articulatory model of the tongue based on a statistical analysis," J. of Acous. Soc. Am., 65:S22, 1979.

[11] Vorperian, H., Kent, R., Lindstrom, M., Kalina, C., Gentry, L., and Yandell, B., "Development of vocal tract length during early childhood: A magnetic resonance imaging study," J. of Acous. Soc. Am., 117:338, 2005.

[12] Kohonen, T., "The self-organizing map," Proceedings of the IEEE, 78(9):1464–1480, 1990.