Speech imitation with a child's voice: addressing the correspondence problem

Miguel Vaz, Holger Brandl, Frank Joublin, Christian Goerick

2009

Speech imitation with a child's voice: addressing the correspondence problem

Miguel Vaz[1,2], *Holger Brandl*[1,3], *Frank Joublin*[1], *Christian Goerick*[1]

[1]Honda Research Institute - Europe GmbH, Offenbach am Main, Germany
[2]Department of Industrial Electronics, University of Minho, Portugal
[3]Research Institute for Cognition and Robotics, University of Bielefeld, Germany
mvaz@dei.uminho.pt

Abstract

We hereby present our first steps towards linking an embodied speech acquisition system and a speech production module, in order to provide a robot with the ability to reproduce acquired speech representations. Due to the type of interaction intended for the robot, we endowed it with a child-like voice, concretized with the use of a new vocoder-like technique for speech synthesis. The task in hand consists of finding and using a correspondence between configurations in the tutor's acoustic parameter space, which might be inaccessible for the system's voice, and phonetically equivalents in the robot's. This mapping is learned by having a cooperative tutor imitating the robot's monophonic utterances, giving the robot the necessary knowledge to map a tutor's utterance to its own vocal space, and imitate it.

1. Introduction

The ability to use natural spoken language to interact with a robotic system like Honda's ASIMO is highly desirable because it greatly increases the naturalness and efficiency of communication. Ideally, in order to face the demand for flexibility of such a task, an acquisition process would make little or no assumptions about the used language and adapt itself to the characteristics relevant in the environment, much like children do in the first years of their lives. The speech representations should be learned through interaction with a tutor instead of being predefined. Traditional ASR and TTS systems rely on annotated corpora, which isn't suitable for our type of interaction scenario.

The first steps towards a system that fulfills the aforementioned requirements have been reported in [1]. There, it has been shown how a system can learn to recognize phones, syllables and words in an unsupervised fashion using child-directed speech. These speech recognition capabilities have already been integrated in an autonomous learning interaction framework working on the humanoid robot ASIMO, where it could acquire speech labels for objects and their attributes, like size, position and orientation [2]. Thanks to these previous works, ASIMO is able to recognize previously learned speech-labels.

It was, however, not yet possible for it to produce an audible description of a perceived scene.We hereby attempt to fill this gap by combining the already mentioned acquisition system with a speech production module.

In order to match its size, it was decided that ASIMO should have a child's voice. A child's voice is also arguably more appropriate for the type of interaction that takes place with a learning system like the one in [2], which has the dimensions of a child and where almost no knowledge of the world is assumed. Synthesizing high-pitched voices, as children's, is however not trivial. There have been recent developments in this direction in the field of articulatory synthesis [3], but synthesizing some consonants, like fricatives, is still a challenge. Acoustic-domain techniques, like those used in state-of-the-art TTS systems [4], also show some limitations in the synthesis high-pitched voices. Because of these limitations we developed a new acoustic-domain technique based on the channel vocoder [5] and using a gammatone filter bank [6]. It allows for the synthesis of high- and low-pitched voices with similar naturalness and intelligibility.

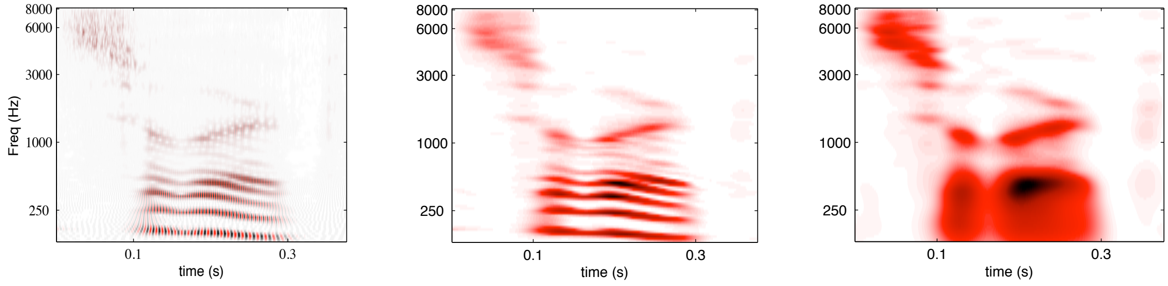
Working with a child-like voice and an adult tutor enables us to address the correspondence problem, because acoustic targets cannot simply be copied from the tutor for reproduction. This is one of the unexplained skills shown by infants during the process of language acquisition: transfer the relevant perceptual auditory features of the utterances of their caregivers into acoustic goals that are attainable by their own different vocal tract. We address this problem by exploring the role of the caregiver's imitative feedback, which is in line with recent views, e.g. [7], of the role of parental imitation in the early speech acquisition process.

In a training phase, the tutor imitates a set of constant spectrum vowel utterances produced by the system. This way, the system associates its vocal productions with the imitative vocal response of the tutor. Later, by inverting this association, the system is able to interpret a tutor utterance as a trajectory in its own articulatory space. As tutor utterances are represented in reference to its vocal repertoire, the system makes no assumptions about the language of interaction. As an initial step, we limited the robot's repertoire to vowels, but because of the frame-wise implementation more complex utterances can be imitated. This way we are able to extend our previous parrot-like real-time speech imitation system [8].

We begin by motivating and describing the speech synthesis framework, which includes a speech synthesis algorithm based on the channel vocoder and an algorithm for morphing the spectral prototypes. We then explain in more detail the imitation mechanism, and describe the experiment done to evaluate its performance. We conclude with a discussion.

2. Speech synthesis framework

Most work done on speech imitation, for example [9] and [10], makes use of articulatory production models. In order to equip our robot with the ability to speak with a child's voice, we use an acoustic production model derived from the channel VOCODER [5] instead. It uses a gammatone filter bank in the ERB scale [6], which offers an optimal tradeoff between spectral and temporal resolution, enabling us to synthesize high- (women's and children's) and low- (men's) pitch voices with similar naturalness and intelligibility.



(a) (b) (c)

Figure 1: Results of the different preprocessing stages involved in the extraction of the spectral features of the German word ”zurück”, from a male speaker. (a) cochlear response, (b) spectral envelope, (c) enhanced spectrogram after filtering over the channels.

Using an acoustic technique for speech imitation offers some advantages, while presenting some difficulties. The first advantage is the ability to synthesize out-of-the-box a broad range of voices, namely children's. Another advantage is the simplicity of control, when compared to an articulatory model: because it is done in a level nearer to the acoustic output, the mapping between target sound and control parameters (inverse mapping) is simpler.

The difficulty posed by an acoustic production model lies in defining a voice model: a set of sounds that can be perceived as having been produced by the same person. In other words, defining a plausible set of acoustic constraints and variability. In standard text-to-speech systems, these are statistically derived from a corpus of utterances [4]. This is however not compatible with our already mentioned goal of learning them through interaction. In this work the voice constraints are implicitly defined by a set of spectral vectors extracted from utterances from a child speaker. The variability is achieved by means of a morphing algorithm that interpolates between these spectral prototypes, using the formant frequencies as correspondence points. This reflects our assumption that, since each of the spectral vectors belongs to the speaker's voice, so do the intermediary steps of a transition between each of these vectors.

We begin by explaining the mechanism of spectral extraction used to compute the spectral prototypes. We then describe the synthesis algorithm that uses these features, and finally the spectral morphing algorithm, used to compute new spectral configurations and transitions.

2.1. Spectral features

The spectral features used to define the system's voice are computed by applying a gammatone filter bank in the ERB scale to an input (child) speech signal. There it is decomposed in several channels (figure 1a). In each of these channels, we compute the Hilbert envelope (figure 1b). The harmonic structure is then removed with an anisotropic gaussian kernel with a width dependent on the pitch value, applied over the channels. Because of this, the formant structure of the resulting spectrogram is more evident (figure 1c). These three processing stages are schematized in figure 2.

2.2. Synthesis algorithm

The synthesis algorithm is based on the channel VOCODER [5]. The excitation signal consists of a sum of sinusoidal functions h_i, fundamental frequency and its harmonics, weighted with a value obtained of directly sampling the spectrogram S at the corresponding frequency, as shown in equations 1 and 2. Fricative excitation, not used in this work, is obtained using the classic VOCODER architecture, by driving white noise through the filter bank and multiplying it with the spectrogram.

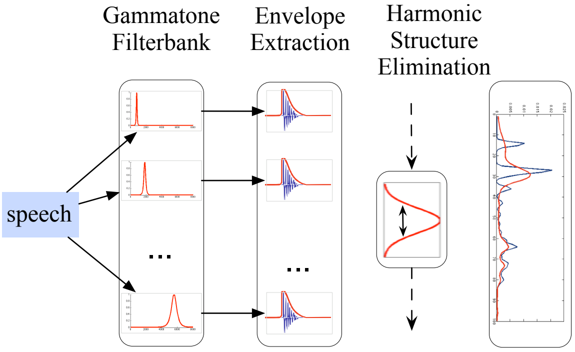


Figure 2: Spectral extraction with the Gammatone filter bank. Each channel is rectified and the Hilbert envelope is extracted. Then, an anisotropic Gaussian filter with a width dependent on the pitch value is applied, in order to enhance the formant structure and eliminate the harmonic structure, and consequently the dependency on the value of the fundamental frequency.

$$h_i(t) = S(c_i(t), t) \ \cos\left(\frac{2\pi}{f_s}\int_0^t f_0(x)dx + \phi_i(0)\right) \tag{1}$$

where $c_i(t)$ is the filter bank channel of the i^{th} harmonic at time t, $\phi_i(0)$ the initial phase of the same harmonic, $f_0(t)$ the fundamental frequency, and f_s the sampling rate.

In addition to this, we add a small weight to each of the harmonics, to compensate the smaller precision of the Gammatone filter bank in the higher frequencies and the additional smoothing effect of the Gaussian kernel, which leads to an overquantification of the energy of the high frequency sinusoidal components. Currently, this consists of a decay of $-0.2dB$ per harmonic, such that the voicing source signal $v(t)$ is given by

$$v(t) = \sum_{i=0}^{n} d_i h_i(t) = \sum_{i=0}^{n} 10^{\frac{-0.2\,i}{10}} h_i(t) \tag{2}$$

An important aspect is the setting of the phase of each of the sinusoidal components $\phi_i(0)$, known to be of high importance to the synthesis quality. The initial phases of the sinusoidal functions are set to linearly distributed, signal alternating, values between $\pi/2$ and 0.

2.3. Spectral morphing algorithm

We use a spectral morphing algorithm for generating the spectral vectors needed for imitating an utterance (see section 3.2). This algorithm has two different functions. One is to compute the transition between two spectral vectors, and the other is to represent an intermediate state between two or more spectral vectors. The algorithm receives initial and final spectral configurations, a set of reference points in both of these vectors, and a correspondence between these reference points. In this work, the correspondence is given by the formant frequencies associated to the spectral vectors. It provides the information needed to connect the two spectral configurations and reconstruct the intermediate steps of the transition.

The morphing algorithm computes the value of spectral channel c at the normalized intermediary position $\alpha \in [0, 1]$, $S(\alpha, c)$, by linearly interpolating between points p_c and q_c, respectively part of the initial and final spectral vectors.

$$S(\alpha, c) = (1 - \alpha)\, S(p_c, 0) + \alpha S(q_c, 1) \tag{3}$$

where p_c and q_c are calculated by maintaining the proportion of the distance from channel c to the immediately inferior, $\overline{p_i q_i}$, and superior, $\overline{p_j q_j}$, line segments as defined by the correspondence matrix

$$\begin{cases} p_c = p_i + \frac{c-c_i}{c_j-c_i}(p_j - p_i) \\ q_c = q_i + \frac{c-c_i}{c_j-c_i}(q_j - q_i) \end{cases} \tag{4}$$

This is illustrated in figure 3.

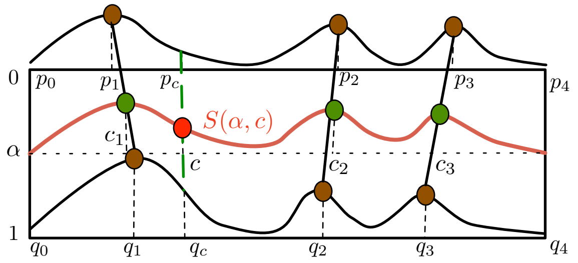


Figure 3: The morphing algorithm computes the spectral value $S(\alpha, c)$ for a channel c at an intermediary position $\alpha \in [0, 1]$, given an initial and final spectral vectors p and q, and a correspondence matrix associating (p_1, q_1), (p_2, q_2), (p_3, q_3).

3. Learning and using a tutor-system mapping for imitation

In some existing models of early infant speech acquisition, the system learns an acoustic to articulatory inverse mapping by direct imitation of its caregiver. However, vocal tracts of children aren't just a shorter version of those of adults, but present significant differences [3]. This reflects in differences in the acoustical characteristics of children's and adult's voices, which include higher fundamental and formant frequencies [11].

In order to make use of the important feedback from the caregiver, the child needs to be able to correspond acoustic targets, suggested by the caregiver, which are unreachable for itself, into other phonetically equivalent targets that he can produce. Depending on the models, the required phonological knowledge is presented as innate [12] or learned [13]. We support an alternative viewpoint, see [7] and [9], that it is the caregiver that has access to this necessary knowledge, which he uses it to interpret the child's immature utterances and imitate them.

In [7], a reinforcement learning model is proposed where the "child" utters in order to receive a reward, usually in the form of a positive response/imitation from the caregiver. This way it learns simple associations between its muscular activity and the caregiver's acoustic output, leaving the complex acoustic voice matching task to the only expert in the linguistic environment: the caregiver. In [9], it is shown how the caregiver can, thanks to its phonological bias, guide the infant in the search for clearer vowel targets. In this work, we investigate the use of the systematic imitation of a caregiver in order to learn a correspondence between the caregiver's and the system's voice, making no assumptions on the tutoring language. In a training phase, the tutor is instructed to imitate utterances from the system. These consist, at this stage, of a set of vowels, synthesized from a set of predefined spectral vectors also seen as motor primitives. The imitative responses of the tutor allow to bootstrap a correspondence model between the system voice and the tutor's one.

In a second stage, given a target tutor utterance, the system computes the posterior probability of each of its motor primitives, based on the previous tutor feedback. These posterior probabilities are then transformed in a time-varying population code of vocal primitives. By means of the morphing operation presented in section 2.3, the population code is transformed into output spectral vectors and synthesized as the system's imitation of the tutor's utterance.

3.1. Training: tutor imitates the system

The system has a repertoire of vocal primitives available in the form of predefined spectral vectors. In the training phase, the task is to learn a correspondence between each of the vocal primitives and the imitation from the tutor. The system produces a set of vowel utterances with constant timbre, by synthesizing each of the spectral vectors. These system utterances are presented to a human tutor, instructed to imitate them. The system analyzes the tutor's responses (see equation 5) and interprets them as imitations to the vocal primitive used to trigger the response. From each utterance, the formant frequencies and their spectral energy are extracted (200Hz). These are then used to compute perceptual features p_i later used in the imitation process:

$$\begin{aligned} p_1(t) &= F_1(t) \\ p_2(t) &= F_2(t) - F_1(t) \\ p_3(t) &= F_3(t) - F_1(t) \\ p_{\{4,5,6\}}(t) &= log(S(C_{\{1,2,3\}}(t), t)) \end{aligned} \tag{5}$$

where $C_i(t)$ is the filter bank channel (ERB scale) of the i^{th} formant frequency F_i, and S the spectral representation obtained by the process described in section 2.1. Feature vectors are then subject to whitening, so that they have mean 0 and variance 1.

3.2. Imitation: system imitates the tutor

In order to imitate, the system maps the perceptual features of the tutor's utterance to an equivalent vocal action. For this, we use a local estimation technique called k-Nearest-Neighbours [14]. The choice was motivated by it not making any assumptions on the data distribution of the tutor's responses. For a set of labels or vocal-classes C_j and an input feature vector x, we consider a neighbourhood V of x that contains exactly k points. The posterior probability of class membership depends on the number of training points of class C_j present in V, denoted by

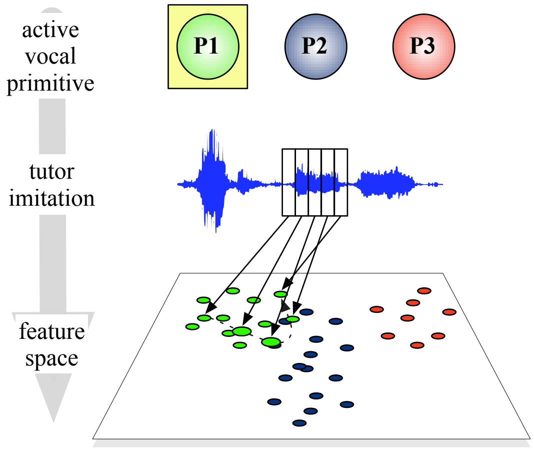


Figure 4: Training procedure for the imitation experiment. A spectral vector is randomly picked (yellow box) from a set of templates and used to synthesize a vowel with constant timbre. A tutor is instructed to imitate the sound that the system produces. The value of the first three formant frequencies and their spectral energy is measured from the tutor utterance and used to train a kNN. The identity of the randomly chosen is used as the label for the perceptual parameters of the tutor's utterance.

K_j:

$$p(C_j|x) = \frac{K_j}{K} \tag{6}$$

Because we work with a small number of basis spectral vectors and want the system to also be able to imitate interphonemic transitions, we need to represent intermediary states. Furthermore, we don't restrict the system's utterances to the basis spectral vectors, in the sense that new sounds can be produced as a combination of the existing primitives. For this, instead of making a hard classification limited to the class C_j that maximizes the posterior probability as given by equation 6, we identify C_{j1} and C_{j2} as the two classes with higher posterior probabilities and use them to code the output spectral vector. We compute the relative strength of activation α of these two classes, given a feature vector x from a tutor utterance, by

$$\alpha = \frac{p(C_{j1}|x)}{p(C_{j1}|x) + p(C_{j2}|x)} \tag{7}$$

This value is used by the morphing algorithm described in 3 for the computation of the spectral output.

The formant frequencies of the new morphed spectral vectors are given by

$$f_i = f_i^{k1}\,\alpha + f_i^{k2}\,(1-\alpha) \tag{8}$$

and can be used as correspondence points in the computation of transitions or new spectral outputs. In order to endow the process with robustness against outliers, class posterior probabilities computed by equation 6 are low-pass filtered in time.

The different stages of the imitation process are shown in figure 6. In figure 6a, we show the spectrogram and the formant frequencies. For each frame, we extract the features and classify them with the k-Nearest-Neighbours algorithm. The classification, expressed in terms of class posterior probabilities, $p(C_j|x)$, is low-pass filtered in time for robustness. The smoothed classification curves, upper part of figure 6b, are used as activations of the motor primitives. These are then morphed accordingly, energy-normalized (lower part of figure 6b), and become the energy contour of the input utterance, and a pitch contour obtained from the original by adding 130Hz.

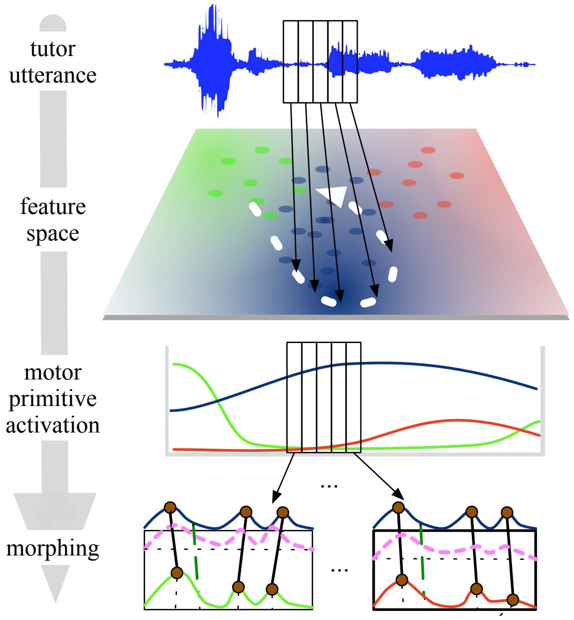


Figure 5: Imitation of a tutor's utterance. The perceptual features of the utterance are analyzed and classified by the kNN algorithm. The posterior probability of each class is computed for each feature vector. The two most likely classes are selected and used as population coding.

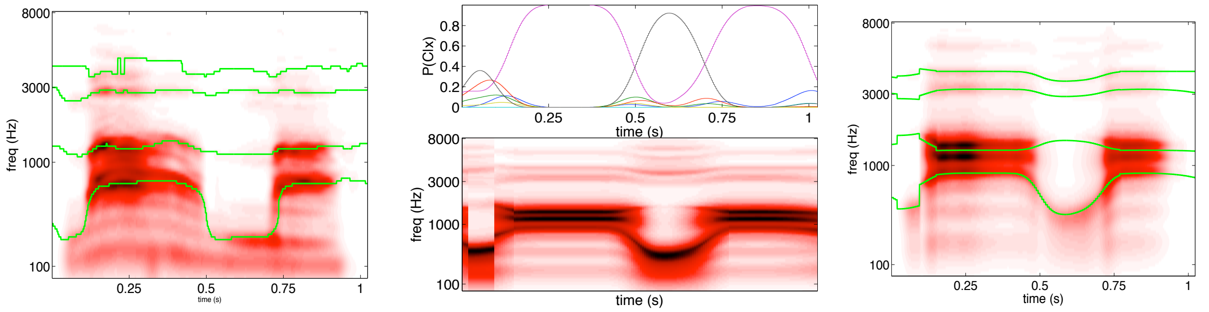
4. Experimental results

We grounded the system's voice in a set of 8 spectral vectors, selected per hand 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. Spectral representations (like in figure 1) were extracted using the procedure described in section 2.1 and the pitch algorithm presented in [16].

The first four formant frequencies were extracted as described in [17], and added to the spectral vectors to generate the correspondence matrices for the morphing algorithm.

Each of the vowel prototypes was synthesized and played 15 times to a male adult speaker, who imitated them. In order to improve the naturalness of the training session we synthesized each robot's utterance with a random duration (between 0.25 to 0.3 seconds), smoothly decaying energy contour and either falling, ascending or flat pitch contours. From each temporal frame (extracted at 200Hz) of each imitative tutor utterance, we extracted the features from equation 5. For the k-Nearest-Neighbours algorithm we used $k = 21$ and the standard euclidean metric.

To evaluate the classification performance of the kNN and the feature quality, we divided the 9355 feature frames into a training and a test sets, containing respectively 66% and 34% of the total number of feature frames. For each class C_j from the test set we computed the mean class posterior probability of



(a) (b) (c)

Figure 6: (a) Spectrum and formant frequencies of the input tutor's utterance [aua]. Upper figure (b): smoothed posterior probabilities of each class/vowel motor primitive. Lower figure (b): energy normalized spectral vectors resulting from the computation of the population coding. Output spectrum (c), with copied envelope.

C_j given a feature frame f

$$M(i,j) = \frac{1}{|test(C_i)|} \sum_{f \in test(C_i)} P(C_j|f) \qquad (9)$$

As can be seen in figure 7, the matrix is mostly diagonal, the sum of the elements of the diagonal corresponding to a ratio 0.7396 of the total sum of the matrix. This shows that the imitative response of the tutor significantly varies depending on the robot's vowel that originated it, suggesting that the system's vowels were clear most of the time.

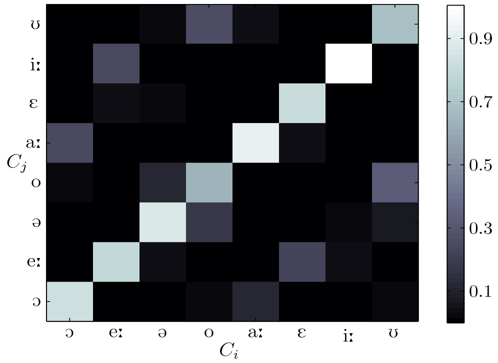


Figure 7: Mean posterior probabilities for each vowel class.

We also performed a subjective evaluation of the imitation process, in order to assess the changes in performance with different vocal repertoires, and the capacity to produce sounds not represented by the vocal primitives. A total of 24 test subjects (17 male and 7 female, all native or experienced german speakers) were presented with 47 pairs of utterances, where the first was taken from a test corpus, with 3 times 13 german vowel utterances spoken by the same male person with which the system was trained, and the second corresponded to an imitation. We asked the participants to judge between 1 and 5 the phonetical similarity of each pair, where 1 corresponds to *different* and 5 to *same*. Used vowels were the 8 represented by the system, and ʏ, ɘ, aɪ, aʊ, ɔɪ.

Imitation utterances were generated using 4 different procedures. We trained the imitation systems with different vocal primitives sets and corresponding training data: the whole set S_8; a shorter set S_3 corresponding to phonemes a, i, ʊ; an intermediate set S_5 containing a, i, ʊ, ɛ, ɔ.The remaining set of utterances was produced by controlled activation of the primitive corresponding to the phoneme being imitated S_c. For the imitation utterances, duration and energy contours were copied from the original, pitch changed by adding 130Hz.

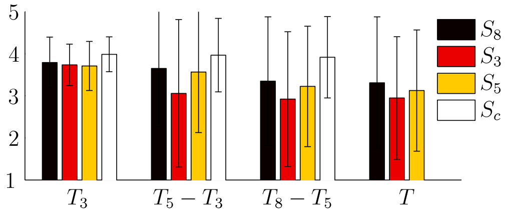


Figure 8: Histogram showing the mean score and variance for all trained (S_3, S_5 and S_8) and control (S_c) systems and different sets of test utterances. T_3 contains a, i, ʊ; $T5 - T_3$ contains ɛ, ɔ; $T_8 - T_5$ contains e, ə, o; and T all 13 phonemes.

The mean scores for each system and test sets are shown in figure 8. As can be observed, the control system S_c performs consistently significantly better than all others. Also generally, through two Kruskal-Wallis (0.05 significance) tests we could find evidence to reject the null-hypothesis that there are neither significant differences between non-control systems nor between test sets. The data also suggest a dependency between the scores of the evaluated phonemes and their representation in the system's repertoire. Namely, statistical significant difference (through a Wilcoxon test at 0.05 significance) can be observed for the performance S_3 and S_5 between their fully represented test sets (T_3 and T_5) and the whole test set. This was not observed for S_8. Furthermore, significant difference was found for T_8 between S_8 and the other systems.

These results confirmed our hypothesis, that the system benefits from a more extended vocal repertoire. Possibly, the ideal case would be that system would have at least one prototype for each of the target categories. Nevertheless, even whilst being directly represented by the system, differences in the scores of the vowels could be observed. This could be due to the clarity of the vowel prototypes themselves, but there is no significant difference shown by the control group. It is then more likely that these are either due to problems with the classification of the tutor utterances, or that through the morphing operation the clarity of the output is affected.

5. Conclusion and future work

We have presented a method to learn a correspondence between vowel sounds from a tutor and vocal actions from its repertoire. The method makes no assumptions on the language of interaction or on the characteristics of the tutor's voice. The mapping is learned by having a cooperative tutor imitating utterances spoken by the system. By analyzing the responses of the tutor, the system first learns to predict the tutor's imitation, and later use it to compute posterior probabilities for each vocal primitive. Using these probabilities, it can represent tutor utterances in terms of its own vocal primitives, and imitate them.

This interpretation of the role of feedback is in tune with recent work in robotics and psychology. It relieves the learning system from the burden of having (to find) a universal speaker-independent speech representation, where its utterances and its tutor's would be similar. Instead, it allows the system to infer from the interaction, which vocal tract configurations produce phonetically equivalent vocal sounds to those of the tutor.

As an initial step, we endowed the system with a predefined set of motor primitives, in the form of spectral vectors corresponding to vowels. We used a nearest-neighbours algorithm to compute the posterior probabilities of the spectral primitives. The system imitates an incoming utterance by transforming the class posterior probabilities into motor primitive activations, which it then merges using a morphing algorithm. The system has been implemented in real-time, and can imitate any vocal sequence, although its repertoire is, for the time being limited to vowels.

We have also described a new speech synthesis algorithm that allows the synthesis of high-pitched voices, without the need of building a specific articulatory model for the desired voice. It is based on the channel-vocoder and uses a gammatone filter bank at its core, which offers an optimal tradeoff between spectral and temporal resolutions.

Presently, the imitation module receives a tutor target utterance from the system, which it then transforms and synthesizes. In the future, however, we will be integrating our system more tightly with the recognition system described in [1], so that we can integrate the recognition and production learning processes of speech acquisition. Furthermore, this will allow us to make use of more sophisticated pattern recognition methods for phone classification, trained on a broader range of utterance examples. This way we will also be able to work with consonant sounds and, more importantly, have a less rigidly divided training and imitation stages.

6. Supplementary information

A supplement to this paper containing the .wav format audio samples with imitation examples can be found on the website: http://mvaz.net/research/specom2009.html.

7. Acknowledgements

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

8. 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] C. Goerick, B. Bolder, H. Janßen, and M. Gienger, "Towards incremental hierarchical behavior generation for humanoids," *IEEE-RAS Int. Conf. on Humanoids*, 2008.

[3] L. J. Boë, G. Captier, J. Granat, M. J. Deshayes, J. Heim, P. Birkholz, P. Badin, N. Kielwasser, and T. Sawallis, "Skull and vocal tract growth from fetus to 2 years," *8th Int. Seminar on Speech Production, Strasbourg*, pp. 157–160, 2008.

[4] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *6th ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.

[5] H. Dudley, "Remaking speech," *Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[6] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer Technical Report #35*, 1993.

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

[8] M. Heckmann, C. Glaeser, M. Vaz, T. Rodemann, F. Joublin, and C. Goerick, "Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis," *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2008*, pp. 1699–1704, 2008.

[9] 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.

[10] 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.

[11] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 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] I. S. Howard and M. A. Huckvale, "Training a vocal tract synthesizer to imitate speech using distal supervised learning," *XI Int. Conf. Speech and Computer*, 2005.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, August 2006.

[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.

[16] M. Heckmann, F. Joublin, and C. Goerick, "Combining rate and place information for robust pitch extraction," in *Interspeech 2007*. ISCA, 2007, pp. 2765–2768.

[17] C. Glaeser, M. Heckmann, F. Joublin, C. Goerick, and H. Grob, "Joint estimation of formant trajectories via spectro-temporal smoothing and bayesian techniques," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 4, pp. 477–480, 2007.