Mapping the vocal tract with a 2D vocalic articulatory space: applications to developmental robotics

Cláudia Soares, Alexandre Bernardino *

Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa, Portugal
csoares@isr.ist.utl.pt, alex@isr.ist.utl.pt

Abstract

Articulatory speech synthesis has been used recently to emulate in robots the speech production and learning capabilities of human infants. Acoustic to motor maps are created by babbling strategies, exploring the available motor degrees of freedom and creating associations to the listened sounds. However, the physiology of the human vocal tract contains many redundant parameters, which poses problems in sensor-motor map learning. In this paper we show that vocalic speech requires, in fact, a very reduced number of parameters and, based on linguistic knowledge, propose a two-dimensional articulatory space. The proposed space is generated through the convex combination of prototype vowels representing extremal points in the articulatory parameters. We show experimentally, using a known articulatory synthesizer, that the proposed model production space is enough to generate most of the vowel acoustic subspace, in terms of the Mel Cepstral Coefficients' variance. This provides a low-dimensional and intuitive vowel production space, suited for automatic production, recognition and learning of speech in articulatory models.

Introduction

Developmental robotics aims at studying how knowledge on human cognitive development can be exploited to allow robot to learn and adapt continuously to its morphology and environment (Lungarella et al., 2003). The development of speech production involves the exploration of the vocal tract capabilities during the infants early developmental stages. Also for speech perception development, the vocal tracts articulatory information may be of fundamental importance. The Motor Theory of Speech Perception (Liberman and Mattingly, 1985) supports that the basic units of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements of the articulator. According to this theory, speech would be perceived by inferring the articulatory shape of the vocal tract from the acoustic signal, and performing recognition in the motor space. The rationale for this approach comes from the fact that motor commands, on the contrary of acoustic signals, are invariant to the environmental conditions, thus providing stable references for recognition.

*This work was supported by EU NEST Project 5010 - Contact, and by Fundação para a Ciência e a Tecnologia (ISR/IST plurianual funding) through the POS_Conhecimento Program that includes FEDER funds.

Initial experimental evidence for the importance of motor information in recognition tasks started with neurophysiological recordings in neurons of the pre-motor cortex of primates, which led to the discovery of Mirror neurons (Gallese et al., 1996). These neurons show spiking activity both when the monkey executes and observes a grasping movement. An experimental study, with a robotic artifact, for the recognition of grasping gestures (Lopes and Santos-Victor, 2005), showed drastic improvements when recognition was based on the motor space rather than the visual space. Mirror neurons are located in the ventral premotor cortex, possibly the homologue of Broca's area in humans, which led to the speculation that action recognition and language production share a common system. Neuroimaging studies of the Broca's region have recently supported this hypothesis in a joint action recognition, language production and grasping task (Hamzei et al., 2003).

Such a theory represents a novel paradigm for speech perception but poses novel challenges since it requires the availability of the agents' motor signals and learning mechanisms for associating the motor and auditory spaces. This can be achieved by exploratory learning (spanning the agent's motor space and observing the outcome in auditory terms), or by imitative learning (listening to other agent's produced sounds and trying to imitate). But, depending on the dimensionality on the involved spaces, this may be too complex to do in practice. A recent model for the control of speech production in humans, the Diva Model (Guenther et al., 2006), follows the motor theory paradigm and accounts for a wide range of acoustic, kinematic and neuroimaging human data. Sensory-motor association is done locally by computing the tangent spaces to the synthesis function at some prototypical points. Mapping the whole articulatory space would require a lot of exploratory learning which, in high dimension spaces, becomes impractical.

In this paper we propose a methodology to create an articulatory subspace in vowel production, allowing a complete characterization of the speech synthesis function and its properties, permitting an feasible online speech processing and learning for robots as Chico and Chica (depicted in Figure 1). The method is motivated by results of Linguistics and Phonetics, where the vowel space is represented in motor terms in a 2D representation. We

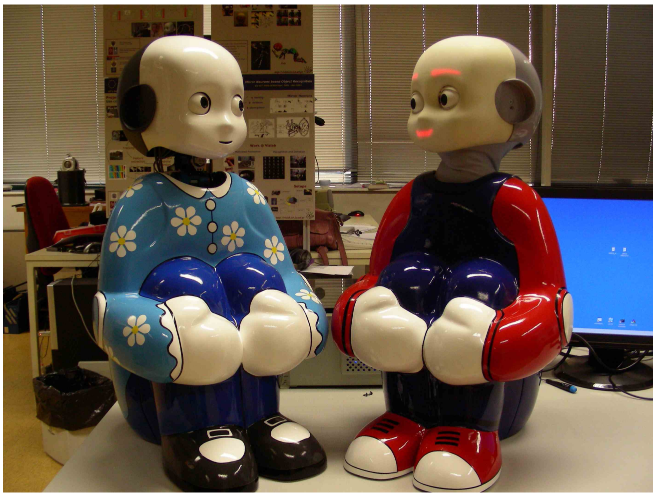
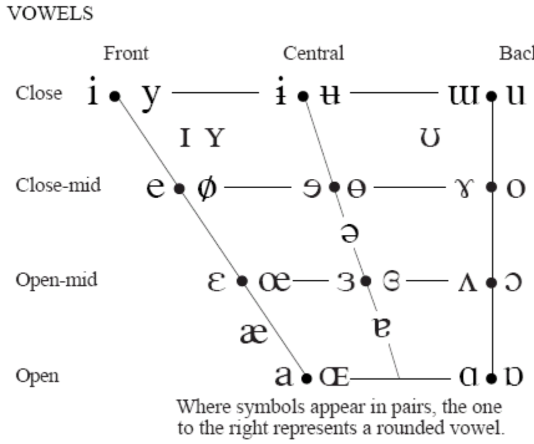
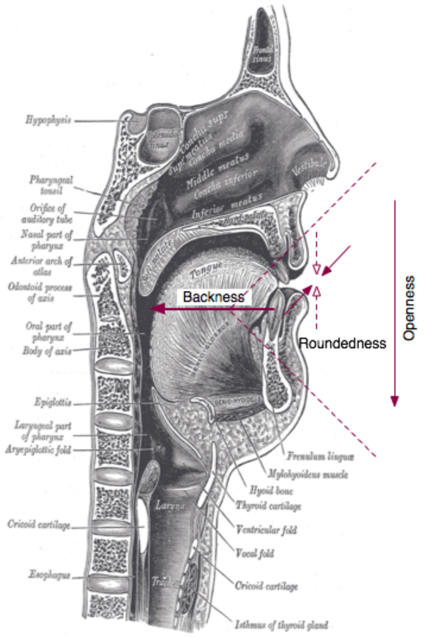


Figure 1: Humanoid robots must interact with each other and with humans by spoken language. These are the robotic platforms for the implementation of the algorithm. Some work has been already done with these robots in speech perception (*vide* (Hörnstein and Santos-Victor, 2007))



(a) International Phonetic Alphabet chart for oral vowels.



(b) Main degrees of freedom represented in the IPA chart. Figure from (Gray, 1918), with our labels.

Figure 2: Articulatory degrees of freedom in the IPA chart representation.

show that a 2-dimensional plane generated by the convex combination of 3 extremal motor primitives is able to adequately represent the vowel acoustic space. An additional advantage is that, since the synthesis is based on two sole articulatory parameters, it is easy and intuitive to graphically visualize the motor-to-acoustic manifold, allowing a better characterization of its properties.

The paper is organized as follows. Section *Linguistic Motivation* briefly presents the linguistics and phonetics results motivating our approach. Then, in Section *The Speech Production Model* we describe the articulatory speech synthesizer used in this work and mathematically formulate the proposed articulatory dimensionality reduction principle. We have performed several experiments illustrating the validity of the approach, presented in Section *Experimental Results*. Finally, Section *Conclusions* present some conclusions and directions for future research.

Linguistic Motivation

Since the beginning of Linguistics and Phonetics speech sounds are classified mainly by articulatory parameters. One of the pioneer works in defining where are vowels located in the articulatory space was (Jones, 1917) in which the mathematician and phonetician Daniel Jones first proposed the Cardinal Vowel Diagram. This diagram was subject of discussion and contributions from the phonetics community and gave rise to the unanimously accepted representation for oral vowels today.

The schematic in the International Phonetic Alphabet (IPA) for oral vowels in Figure 2(a) shows the distribution of vocalic sounds in three dimensions relative to the human vocal tract: height (vertical axis), backness (horizontal axis) and roundedness (lip rounding)(Association, 1999) as illustrated in Figure 2(b).

This choice of reference frame has roots in the physiology of the phonatory system. The vocal tract configuration for oral vowels is function of the tongue, the jaw and the lips. The jaw and lips can have several degrees of openness, the tongue can assume the articulatory positions in front, center or back of the oral cavity and the lips can also change the vocal tract by rounding. So, these three articulatory parameters are considered the main degrees of freedom of vocalic speech sounds, and represent the directions that better explain the inter-vowel variation. Nevertheless, there are other static articulatory parameters that influence oral vowel quality, although they are not determinant in most spoken languages.

In most languages, rounded and unrounded vowels are not minimal pairs, i.e., for the same articulatory configuration, roundedness alone does not create two different phonological vowels. In addition to this, some studies support that roundedness is perceived mainly by vision in normal hearing-seeing subjects (Traunmüller, 2006), For these reasons, the main articulatory dimensions considered for oral vocalic sounds in the human vocal tract are the height and backness, motivating the approximation proposed in this paper — whatever the dimensionality of the articulatory space we consider, there is an two-dimensional subspace approximation that maps the vowel system of most languages. The phones [i], [a] and [u] define a set of axis in the 2D plane of the articulatory parameters of *height* and *backness*. These three vowels are called *corner vowels* because they represent extreme placements of the tongue forming the corners of a triangle in articulatory space. They also form a triangle in formant space (F1 – F2)(Titze, 1994). Therefore, we consider these phones the extremal points in our model, and will produce the remaining ones by their convex combination. This will be detailed in the following Section.

The Speech Production Model

To test and validate our proposal we use a well-known articulatory speech synthesizer. This will allow us to do systematic tests and quantify the errors arising from the proposed approximation. From realizations of the extremal phones [i], [a] and [u], we generate a dense representation of the feasible acoustic signals. Then, to evaluate the model, we compute the acoustic errors outside the feasible set.

Articulatory Synthesizer

The synthesizer in use[1] is a Matlab version of Shinji Maeda's Vocal Tract Calculator (*VTCalcs*) (Maeda, 1990). The seven articulatory parameters are *jaw, tongue, shape, apex, lip_ht* (lip hight), *lip_pr*, (lip protrusion), *larynx*. Each one can assume any value in $[-3;3]$. The articulator parameters are presumed independent, which is not the case in the human vocal tract, leading sometimes to improbable configurations of the articulators, producing a non human sound or even no sound at all. In fact, after a dense sampling of the six-dimensional hypercube and feeding the samples to the synthesizer, as explained later in this section, we realized that only 44.22% of the articulatory vectors generated sound, even if not a human-like one.

The space of the articulators in *VTCalcs* is homographic to $\mathbb{R}^7$, but to produce vocalic voiced sounds only 6 parameters are distinctive, since larynx controls the voicing.

The synthesizer's output is a sound represented by its temporal amplitude. To analyze the sound waveform we use the Mel Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein, 1980), using 12 coefficients.

[1]Available at the CNS Speech Lab webpage `http://speechlab.bu.edu/VTCalcs.php`

Let vector $\mathbf{v} \in \mathcal{V} \subset \mathbb{R}^6$ represent a configuration of the six-dimensional synthesizer's articulatory space and $\mathbf{a} \in \mathcal{A} \subset \mathbb{R}^{12}$ be a vector of MFCC coefficients in the acoustic space. We define the synthesis function as:

$$f : \mathcal{V} \mapsto \mathcal{A}, \quad \mathbf{a} = f(\mathbf{v}) \tag{1}$$

The function is not invertible — distinct articulatory configurations may lead to very similar sounds (in particular, many configurations generate no sound at all). Therefore, there is ambiguity in the identification of motor configurations corresponding to the listened acoustic signals, which may pose problems to motor-based learning and recognition algorithms. To deal with this we define a subspace of $\mathcal{V}$ where the restriction of f to this subspace is assumed invertible.

Dimensionality Reduction

We define a two-dimensional subspace of the full articulatory space, generated by a convex combination of vowels corresponding to extremal positions in the articulatory space. There are two major arguments that support this approach: a linguistic argument, and a experimental one. As mentioned in Section *Linguistic Motivation*, according to Linguistics and Phonetics knowledge, most of the vowel production capabilities of the human vocal tract can be explained by two parameters related to the height and backness of the articulators. The experimental argument is that the Isomap, as discussed in Section *Experimental Results*, shows that there is a good two dimensional approximation to the image of f.

Considering the $\mathbb{R}^6$ prototypes for the extremal phones [i],[a] and [u], it is possible to generate an affine space with all the properties of a convex space. Let a_0, u_0 and $i_0 \in \mathbb{R}^6$ be the chosen vowel prototypes for [i], [u] and [a] and a two-dimensional vector $\mathbf{p} \in \mathcal{V} : \mathbf{p} = (\alpha, \beta)$, with α and β real parameters. A convex combination of the given points forming a 2-dimensional triangle, can be defined by the function:

$$v : \mathcal{P} \subset \mathbb{R}^2 \mapsto \mathcal{M} \subset \mathcal{V}$$
$$v(\alpha, \beta) = \alpha\, i_0 + \beta\, a_0 + (1 - \alpha - \beta)\, u_0$$

where the input space $\mathcal{P}$ is defined as:

$$\mathcal{P} = \{(\alpha, \beta) : \alpha + \beta \leq 1 \wedge \alpha, \beta \geq 0\}$$

Let $\mathcal{M}$ be the image of v, and denote it the *Motor Space*. We define the function f_2 as the restriction of the synthesizer's function f to the motor space, and call it's image $\mathcal{A}_2$

$$f_2 : \mathcal{M} \mapsto \mathcal{A}_2 \subset \mathcal{A}. \tag{2}$$

We will denote f_2 as the *Motor-Acoustic Map*. The image of this function will produce a 2D manifold $\mathcal{A}_2$ in the MFCC acoustic space. Given the choice of the *Motor-Space*, the properties of the used synthesizer (assuming smoothness), and the dense sampling made on $\mathcal{M}$, there are strong reasons to believe that f_2 is invertible. Therefore, the inverse function of f_2, f_2^{-1} is an acoustic to motor

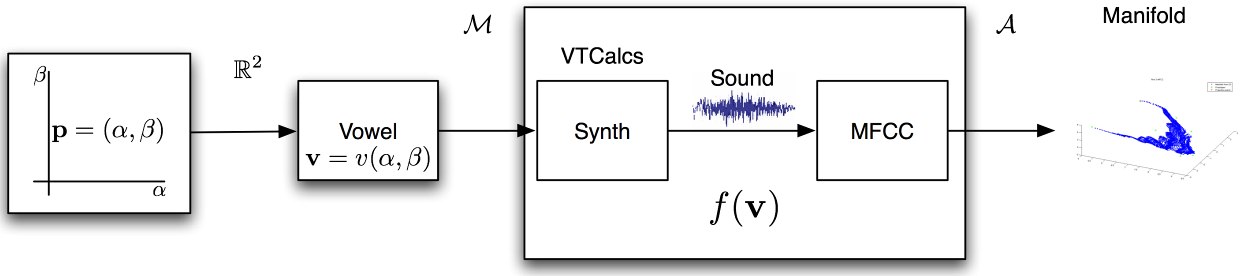


Figure 3: Vowel generation diagram.

map. A schematic representation of the proposed vowel production model is shown in Figure 3.

The twelve-dimensional acoustic space was sampled twice; one representing the span of the reduced articulatory space (using the motor map f_2 from the motor space $\mathcal{M}$), and another representing the span of the full articulatory space (from $\mathcal{V}$). We will show that the former contains most of the information present in the latter.

To estimate the *acoustic manifold* $\mathcal{A}_2$ we have sampled the parameter space $\mathcal{P}$ in steps of 0.01 in the α and β parameters, generating a discrete set of 5000 samples:

$$\mathcal{P}_d = \{\mathbf{p}_i,\ i = 1, \ldots, 5000\}$$

These samples were then used to generate a motor-space sample set, using function v:

$$\mathcal{M}_d = \{\mathbf{m}_i = v(\mathbf{m}_i),\ i = 1, \ldots, 5000\}$$

Thus a discrete sampling of the acoustic manifold was created using the synthesizer's function:

$$\mathcal{A}_{2d} = \{\mathbf{a}_i = f_2(\mathbf{m}_i),\ i = 1, \ldots, 5000\} \tag{3}$$

The first three coordinates of the sampled acoustic manifold are plotted in Figure 4.

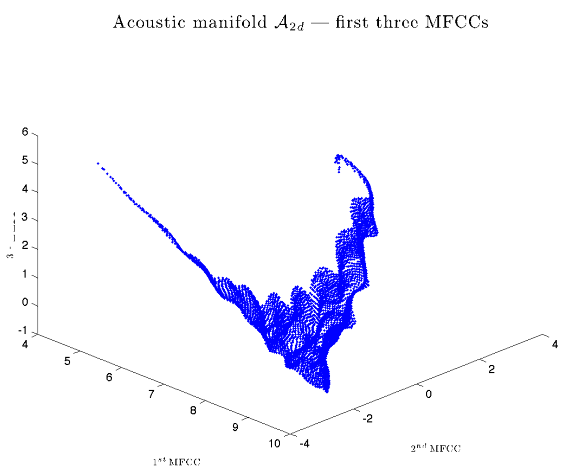


Figure 4: Representation of the first three Mel coefficients of the acoustic manifold.

The *VTCalcs* parameter's six-dimensional $\mathcal{V}$ space was also sampled in steps of 0.6 obtaining a grid with 10 samples per dimension. The point cloud has 10^6 samples:

$$\mathcal{V}_d = \left\{\mathbf{v}_i,\ i = 1, \ldots, 10^6\right\}$$

Again, the synthesizer's function was applied to the data;

$$\mathcal{A}_d = \left\{\mathbf{a}_i = f(\mathbf{v}_i),\ i = 1, \ldots, 10^6\right\} \tag{4}$$

From this data it was removed the set of samples with zero sound amplitude, retaining 44.22% of the initial number.

Experimental Results

To validate the proposed model we generate a set of test vowels $\mathbf{a}^t$ and compute the error in acoustic space (MFCC coefficients) between each one and its projection on the manifold $\mathcal{A}_{2d}$. We also consider the residual variance incurred in a two dimensional approximation of $\mathcal{A}$.

Since we do not have a analytic expression for the $\mathcal{A}_2$ surface, we use its sampled version defined by equation (3). To compute the projection of each point we use the nearest neighbor operator:

$$nn(\mathbf{a}^t) = \left\{\mathbf{a}_i \in \mathcal{A}_{2d} : i = \underset{i}{\operatorname{argmin}} \left\{\|\mathbf{a}_i - \mathbf{a}^t\|_2\right\}\right\} \tag{5}$$

The acoustic approximation error is then computed by:

$$E_a(\mathbf{a}^t) = \|\mathbf{a}^t - nn(\mathbf{a}^t)\|_2 \tag{6}$$

The acoustic approximation error relative to the size of the manifold is defined as

$$\delta_a(\mathbf{a}^t) = \frac{E_a(\mathbf{a}^t)}{\max(length(\mathcal{A}_{2d}))}\ 100\% \tag{7}$$

This measure is dimensionless and gives an indication of how good is the approximation relative to the size of the approximating surface. We consider acceptable to use the maximum length of $\mathcal{A}_{2d}$ to normalize the error because the manifold's shape is not too discrepant, as it is possible to confirm in the Isomap embedding shown in Figure 5. This embedding was determined with the Isomap algorithm

Proc. LangRo'2007
978-972-96895-2-9

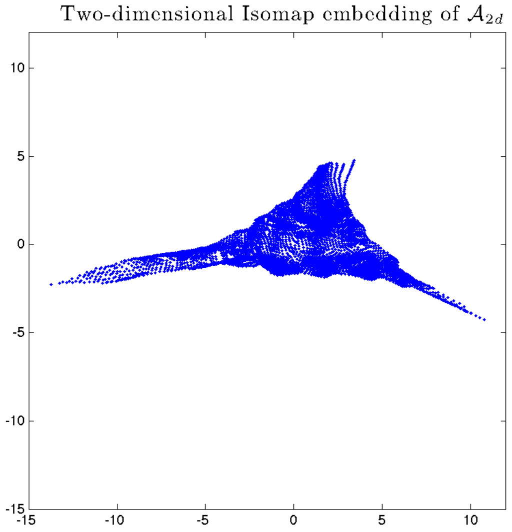


Figure 5: Isomap embedding for the two-dimensional manifold $\mathcal{A}_{2d}$.

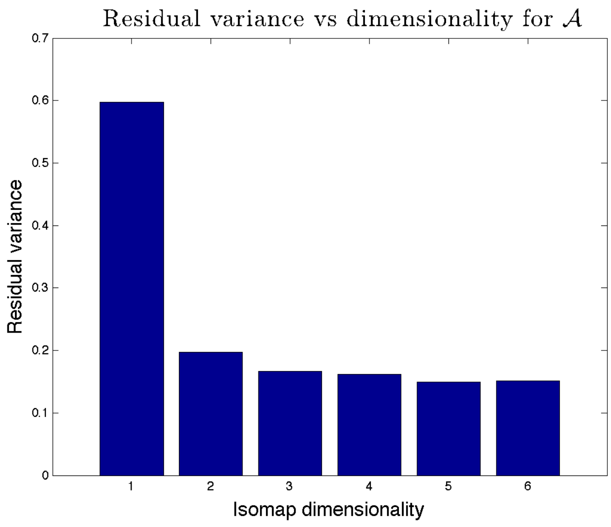


Figure 6: The Isomap algorithm provides the residual variance of the fit to the model's dimensionality. The greatest decrease in variance happens from one to two dimensions of the manifold representing the global acoustic space $\mathcal{A}$.

as described in (Tenenbaum et al., 2000). The *isometric feature mapping procedure* or Isomap recovers low-dimensional nonlinear structure in perceptual datasets. It finds a space embedding for the data, preserving its intrinsic metrics, by conserving distances measured through *geodesic paths* along the observation manifold. For $\mathcal{A}_{2d}$, Isomap created a reproduction, in the two-dimensional space, of the pairwise distances measured in the acoustic twelve-dimensional space.

Dimensionality reduction: validation

To validate the goodness of a two-dimensional approximation for the full space $\mathcal{A}$, the dimensionality of the sampled space $\mathcal{A}_d$, defined in equation (4), was investigated.

Through Isomap we estimate that the dimensionality of the image of f is 2, with a residual variance of 0.197, as illustrated in Figure 6.

The global articulatory space $\mathcal{M}$ is six-dimensional, thus the maximum possible dimensionality for $\mathcal{A}$ is six because f is continuous. The residual variance of the data for six or more dimensions can be interpreted with regard to phenomena such as noise and numerical problems in the MFCCs calculation.

This experimental result confirms that there is a good two dimensional approximation to the overall acoustic space $\mathcal{A}$. The residual variance present in the 2D approximation is partially due to the model simplification but its slow decrease with dimensionality leads to the conclusion that it is caused mainly by non informative phenomena.

Vowel prototypes: appropriateness

To investigate the performance of the approximating space with speech sounds of real languages, some experiments have been conducted with synthesized prototypes of several languages. Those prototypes may lie outside the motor space $\mathcal{M}$ because there are many redundant articulatory configurations that generate the same vocalic sound. We want to show here that $\mathcal{M}$ is complete, i.e. it contains a configuration generating an (almost) identical sound.

Some prototype vowels used in the tests are included in the *VTCalcs* matlab package and are preexistent to the experiment; the other sets were constructed by us and validated by naive native speakers. The speech sounds intensity, fundamental frequency and duration were kept constant so to validate strictly the model for vocal tract configuration.

In the *VTCalcs* package there are eleven prototypes for oral vowels which are found outside the two-dimensional polygon $\mathcal{M}$. They were used to evaluate the amount of error introduced in the two-dimensional approximation. The error was measured as described above, and the results are shown in Table 1. The oral vowels from two very distinct european languages were also used for the same purpose: vowels from Portuguese, an indo-european, romanic language, and vowels from Finnish, a finno-ungric language. Nine Portuguese prototype vowels were used. The errors are shown in Table 2. From Finnish, the eight short vowels were investigated, with results that can be seen in Table 3.

The sample mean over the percent error $\delta_a(\mathbf{a}^t)$ is 2.95% in the portuguese vowels set, 3.87% in the finnish vowels, and 2.23% in the *VTCalcs* set. The standard deviation is 2.22%, 2.81% and 2.02% in the portuguese, finnish and *VTCalcs* sets, respectively. The maximum value for the percent error is 9.17% in the finnish dataset.

So, in terms of the error, the two-dimensional convex space performs well with linguistically relevant synthesized speech sounds. Acoustically, the prototypes and the projections are hardly distinguishable. Inverting the projected points through f_2^{-1} back to the two dimensional motor space $\mathcal{M}$, and plotting the result (Figure 7) makes it

Table 1: *Approximation error for the* VTCalcs *prototypes.*

vowel	symbol	$E_a(\mathbf{a}')$	$\delta_a(\mathbf{a}')\%$
1	iy	0.40149	1.6295
2	ey	0.17829	0.72361
3	eh	0.1522	0.61771
4	ah	0.48633	1.9738
5	aa	0.24348	0.98818
6	ao	0.51035	2.0713
7	oh	0.58974	2.3935
8	uw	1.6111	6.5389
9	iw	1.4057	5.7053
10	ew	0.29547	1.1992
11	oe	0.18119	0.73536

Table 2: *Approximation error for the portuguese prototypes.*

vowel	IPA symbol	$E_a(\mathbf{a}')$	$\delta_a(\mathbf{a}')\%$
1	ɨ	0.13425	0.54487
2	ɐ	1.2335	5.0061
3	ɛ	0.37961	1.5406
4	ɔ	0.50396	2.0453
5	e	0.61689	2.5037
6	o	1.4141	5.739
7	a	0.24161	0.98057
8	u	1.6211	6.5792
9	i	0.39633	1.6085

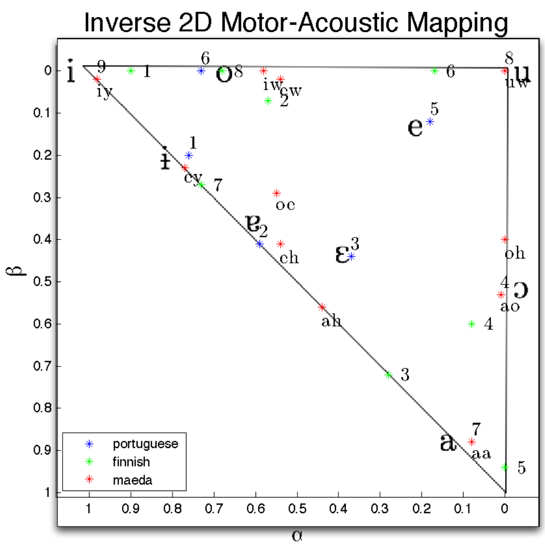


Figure 7: The inverse mapping of the vowel prototypes. The Portuguese vowels are numbered as in Table 2, and the Finnish as in Table 3. Some landmark IPA phonetic symbols are also represented.

is possible to extract some similarities between the IPA openness and backness and the motor space α and β parameters. The hypothesis that the restrictions in the con-

Table 3: *Approximation error for the finnish prototypes.*

vowel	IPA symbol	$E_a(\mathbf{a}')$	$\delta_a(\mathbf{a}')\%$
1	i	0.28764	1.1674
2	ø̞	0.7918	3.2135
3	æ	0.99949	4.0564
4	o̞	0.87593	3.555
5	ɑ	1.6373	6.645
6	u	0.5645	2.291
7	e̞	0.21044	0.85406
8	y	2.2605	9.1741

struction of $\mathcal{M}$ can be used to simulate physiological constraints, is corroborated by these experimental results.

Conclusions

In this paper we have proposed a two dimensional parameterization for the motor space of an available speech synthesizer, *VTCalcs*. The approach is able to generate acoustic signals that represent well all the vowels produced by the synthesizer. Namely, the euclidean error relative to the size of the two dimensional approximating surface has an average of about 3% and a maximum of 9.17% in the used test sets, and the Isomap analysis of the residual variance versus the dimensionality of the approximating manifold confirms the validity of a two-dimensional model for the overall acoustic space.

The proposed model is important by two main reasons:

- The motor space is two-dimensional, thus can be densely sampled with low computational requirements. This simplifies creation and representation of the motor acoustic map.

- The restriction of the synthesizer's function to the proposed motor-space is invertible, allowing to map signals back from the acoustic to motor coordinates.

In future work we will apply the proposed model in the early stages of autonomous speech learning of humanoid robots. The fact that this space has low dimensionality facilitates initial bootstrapping. We will also consider the problem of Mel Coefficients robustness and normalization procedures on the signals.

Since the acoustic manifold appears to be smooth, we will provide it with a differential structure and use it for local optimization, e.g. for guided exploratory learning in imitation tasks. In the long term we intend to apply the proposed model in the early stages of autonomous speech learning of an humanoid robot. The fact that this space has a dimensionality of two, facilitates its bootstrapping role in autonomously to produce and recognize speech. Once the system learns a good initial model of the motor-audio map using the low dimensional manifold, it can expand the available degrees of freedom and refine its production capabilities. As in the ontogenesis of humans infants, such a developmental strategy is more likely to succeed than learning from scratch with the whole system's complexity.

Proc. LangRo'2007
978-972-96895-2-9

References

Association, I. P. (1999). *Handbook of the International Phonetic Association*. CUP, Cambridge.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, 28(4):357–366.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2):593–609.

Gray, H. (1918). *Anatomy of the Human Body*. Lea & Febiger, Philadelphia.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96(3):280–301.

Hamzei, F., Rijntjes, M., Dettmers, C., Glauche, V., Weiller, C., and Buchel, C. (2003). The human action recognition system and its relationship to broca's area: an fmri study. *NeuroImage*, 19(3):637–644.

Hörnstein, J. and Santos-Victor, J. (2007). A unified approach to speech production and recognition based on articulatory motor representations. IROS07, to appear.

Jones, D. (1917). An english pronouncing dictionary. In *Daniel Jones: Selected Works*. Routledge, London.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Lopes, M. and Santos-Victor, J. (2005). Visual learning by imitation with motor representations. *Systems, Man and Cybernetics, Part B, IEEE Transactions on*, 35(3):438–449.

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *Connection Science*, 15(4):151 — 190.

Maeda, S. (1990). *Speech production and speech modeling*, chapter Compensatory articulation during speech, evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model, pages 131 — 149. NATO ASI Series. Kluwer Academic Publisher, Dordrecht, Netherlands.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Titze, I. R. (1994). *Principles of Voice Production*. Prentice Hall, Eaglewood Cliffs, New Jersey.

Traunmüller, H. (2006). Cross-modal interactions in visual as opposed to auditory perception of vowels. In *Proceedings of Fonetik 2006, the XIXth Swedish Phonetics Conference*, pages 137—140. Department of Linguistics, Lund University.