A Unified Model Of The Origins Of Phonemically Coded Syllable Systems

Pierre-yves Oudeyer
Sony Computer Science Lab, Paris
e-mail : py@csl.sony.fr

Abstract

Human sound systems are invariably phonemically coded, which means that there are parts of syllables that are re-used in other syllables. It is one of the most primitive compositional system in language. To explain this phenomenon, there existed so far three kinds of approaches : "Chomskyan"/cognitive innatism, morpho-perceptual innatism and the more recent approach of "language as a complex cultural system which adapts under the pressure of efficient communication". We proposed in (Oudeyer 2002) a new hypothesis based on a low-level model of sensory-motor interactions, characterized by the absence of functional pressure and the use of very generic neural devices. This paper presents a unified model of the origins of syllable systems which does allow a comparison of the different hypothesis on the same ground. We show that our hypothesis is the only one to be sufficient, and that all others are not necesary. Moreover, the model we present the first that shows how a population of agents can build culturally a complex sound systems without the assumption that they already share a phonemic repertoire.

What does explain phonemically coded syllable systems ?

Human sound systems have very particular properties. Perhaps the most basic is that they are phonemically coded. This means that syllables are composed of re-usable parts. These are called phonemes. Thus, syllables of a language may look rather like la, li, na, ni, bla, bli, etc ... than like la, ze, fri, won, etc .... This might seem unavoidable for us who have a phonetic writing alphabet, but in fact our vocal tract allows to produce syllable systems in which each syllable is holistically coded and has no parts which is also used in another syllable. Yet, as opposed to writing systems for which there exists both "phonetic" coding and holistic/pictographic coding (for e.g. Chinese), all human languages are invariably phonemically coded.

The question is then : Why is this so ? How did it appear ? What are the genetic, glossogenetic/cultural, and ontogenetic components of this formation process ? These questions are of particular interest and generality since phonemic coding is a form of primitive compositionality. Compositionality is thought to be the keystone of syntax, and thus understanding how it appeared might help a lot to understand syntactic languages which make humans unique. Several approaches have already been proposed in the literature.

The first one, known as the "post-structuralist" Chomskian view, defends the idea that our genome contains some sort of program which is supposed to grow a language specific neural device (the so-called Language Acquisition Device) which knows a priori all the algebraic structures of language. This concerns all aspects of language, ranging from syntax to phonetics (Chomsky and Halle, 1968). In particular this neural device is supposed to know that syllables are composed of phonemes which are made up by the combination of a few binary features like the nasality or the roundedness. Learning a particular language only amounts to the tuning of a few parameters like the on or off state of these features. It is important to note that in this approach, the innate knowledge is completely cognitive, and no reference to morpho-perceptual properties of the human articulatory and perceptual apparatuses appears. This view is becoming more and more incompatible with neuro-biological findings (which have basically failed to find a LAD), and genetics/embryology which tend to show that the genome can not contain specific and detailed information for the growth of so complex neural devices.

Another approach is that of "morpho-perceptual" innatists. They argue (Stevens 1972) that the properties of human articulatory and perceptual systems explain totally the properties of sound systems. More precisely, their theory relies on the fact that the mapping between the articulatory space and the acoustic and then perceptual spaces is highly non-linear : there are a number of "plateaus" separated by sharp boundaries. Each plateau is supposed to naturally define a category. Hence in this view, phonemic coding and phoneme inventories are direct consequences of the physical properties of the body. Yet, it seems that there are flaws to this view : first of all, it gives a poor account of the great diversity that characterize human languages. All humans have approximately the same articu-

latory/perceptual mapping, and yet different language communities use different systems of categories. One could imagine that it is because some "plateaus"/natural categories are just left unused in certain languages, but perceptual experiments (Kuhl 2000) have shown that very often there are sharp perceptual non-linearities in some part of the sound space for people speaking language L1, corresponding to boundaries in their category system, which are not perceived at all by people speaking another language L2. This means for instance that Japanese speakers cannot hear the difference between the "l" in "lead" and the "r" in "read". As a consequence, it seems that there are no natural categories. This paper will provide quantitative evidence that the morpho-perceptual innatism hypothesis is not a satisfying candidate.

A more recent approach proposes that the phenomena we are interested in come from self-organization processes under functional pressures occurring mainly at the cultural and ontogenetic scale. The basic idea is that sound systems are good solutions to the problem of finding an efficient communicative system given articulatory, perceptual and cognitive constraints. And good solutions are characterized by the regularities that we try to explain, in particular phonemic coding. This approach was initially defended by (Lindblom 1992) who showed for example that if one optimizes the energy of vowel systems as defined by a compromise between articulatory cost and perceptual distinctiveness, one finds systems which are phonemically coded which means that some targets composing syllables are re-used (note that Lindblom presupposes that syllables are sequences of targets, which we will do also in this paper). Yet, these results were obtained with very low-dimensional and discrete spaces, and it remains to be seen if they remain valid when one deals with realistic spaces.

These experiments were a breakthrough as compared to innatists theories, but provide unsatisfying hypothetical explanations : indeed, they rely on explicit optimization procedures, which never occur as such in nature. There are no little scientists in the head of humans which make calculations to find out which vowel system is cheaper. Rather, natural processes adapt and self-organize. Thus, Lindblom's model does not really provide explanations, and one has to find the processes which formed these sound systems, and can be viewed only a posteriori as optimizations. This has been done for the questions of vowel inventories regularities : indeed, in spite of the fact that our vocal tract allows us to produce thousands of different vowels, languages of the world use rarely more than 10 of them, and most often 5 of them. Moreover, among these actually used vowels, some of them appear very often (e.g. [a], [i] or [u] in 89 percent of languages) and others are very rare (e.g. [y]). Lindblom proposed a model which again optimized motor and perceptual constraints, which predicted these regularities. (de Boer 2001) developed an explanatory model in which the basic processes which do produce realistic vowel systems are imitation behaviors among humans/agents. He built a computational model which consisted of a society of agents playing culturally the so-called "imitation game". Agents were given a physical model of the vocal tract, a model of the cochlea, and a simple prototype based cognitive memory. (Oudeyer 2001b) extended this model by letting agents produce complex utterances (syllables), and showed how realistic phonotactic regularities (e.g. the sonority hierarchy principle, the high occurrence of CV syllables, etc...) could emerge without being explicitly programmed in. Yet, as far as phonemic coding is concerned, which is the focus of interest in this paper, (Oudeyer 2001b) does not provide explanations : possible phonemes that agents can use were pre-given and phonemic coding were pre-programmed. There is clearly a need to extend this model by not giving initially a limited and discrete set of possible phonemes and by not coding in phonemic coding. This is what we are going to present in this paper. Interestingly, the solution will rely on a model by (Oudeyer 2002) which was initially developed to explore the last hypothesis of phonemic coding.

Indeed, (Oudeyer 2002) is so far the only truly explanatory model for the phenomenon of phonemic coding. The hypothesis it proposes is that phonemic coding might be a non-functional consequence of sensory-motor coupling. "Non-functional" means that as opposed to models presented in last paragraph, there is no pressure of efficient communication. Phonemic coding, which is indeed useful to develop efficient communication system, yet would not have appeared for this task but may have been recruited only afterwards, being available "by chance" : this kind of phenomenon is sometimes called "exaptationism". The model is at a lower-level than others since it uses neural cortical maps, and their dynamics, coupling perception and action, provides one explanation for phonemic coding. (Oudeyer 2002) explained that this model was not incompatible with functional models such as (de Boer 2000, Oudeyer 2001b), but rather could be a possible manner to bootstrap imitation games. This is what we are going to show in this paper by integrating (Oudeyer 2001b) and (Oudeyer 2002). A problem appeared in previous research when one tried to have agents play imitation games with complex utterances : there were two levels, i.e. the level of articulatory targets/phonemes, and the level of syllables (sequences of phonemes). With simple binary feedback signal, it was difficult to know what kind of errors one agents may have done : wrong number of phonemes ? right number but one is badly imitated ? one phoneme is unknown ? or

is this the new complex sound which is unknown ? In the model presented here, these problems disappear since the low level of targets/phonemes works without supervision. Finally, the model is flexible enough to allow an implementation of all the non-cognitive innatist hypothesis' : morpho-perceptual innatism, functionalism of Lindblom, and exaptationism of Oudeyer.

We will first present an overview of the model in (Oudeyer 2002), then extend it so as to unify it with (Oudeyer 2001b), and finally show how the tuning of some parameters allows to instantiate the various hypothesis concerning the origins of phonemic coding. Then we will present results evaluating each hypothesis.

The coupled neural maps model

The model is based on topological neural maps. This type of neural network has been widely used for many models of cortical maps (Morasso et al., 1998). It relies on two neuroscientific findings (Georgopoulos 1988) : on the one hand, for each neuron/receptive field in the map there exist a stimulus vector to which it responds maximally (and the response decreases when stimuli get further from this vector) ; on the other hand, from the set of activities of all neurons at a given moment one can predict the perceived stimulus or the motor output, by computing what is termed the population vector (see Georgopoulos 1988) : it is the sum of all preferred vectors of the neurons ponderated by their activity. When there are many neurons and the preferred vectors are uniformly spread across the space, the population vector corresponds accurately to the stimulus that gave rise to the activities of neurons, while when the distribution is inhomogeneous, some imprecisions appear. (Oudeyer 2001a) showed that this imprecision allows to explain the well-known phenomenon of "perceptual magnet effect" (Kuhl 2000), which is a perceptual warping of the acoustic space. Moreover, the neural maps are recurrent, and their relaxation consists in iterating the coding/decoding with the population vector : the imprecision coupled with positive feedback loop forming neuron clusters provides well-define non-trivial attractors which can be interpreted as (phonemic) categories.

There are two neural maps : one articulatory which represents the motor space, and one acoustic which represent the perceptual space. The two maps are fully connected to each other with symmetric weights. These weights are supposed to represent the correlation of activity between neurons, and allow to perform the double direction acoustic/articulatory mapping. They are learnt with a hebbian learning rule.

The network is initially made by initializing the preferred vectors of neurons to random vectors following a uniform distribution. Part of the initial state can be visualized by plotting all the preferred vectors as in one of the upper squares of figure 1 which represents the acoustic maps of two agents (the perceptual space is 2-dimensional, and points represents the preferred vectors of neurons). One can also visualize the initial attractors of the acoustic neural maps : the lower squares of figure 1 show examples, in which each arrow has its ending point being the population coded vector after one iteration of the relaxation rule with initial activation of neurons corresponding to the population vector represented as the beginning of the arrow. What one can notice is that initially, attractors are few, trivial and random (most often there is only one).

Then there is a learning mechanism used to update the weights/preferred vectors in the two neural maps when one agent hears a sound stimulus which is represented by a temporal sequence of feature vectors, typically corresponding to the formants of the sound at a moment t (formants are the frequencies for which there is a peak in the power spectrum). For each of these feature vectors, the activation of the neurons in the acoustic map is computed, which propagates to the motor map. Then, each neuron of each map is updated so as to be a little bit more responsive to the perceived input next time it will occur (which means that their preferred vectors are shifted towards the perceived vectors).

The agents in this model produce dynamic articulations. These are generated by choosing N articulatory targets, and then using a control mechanism which drives the articulators successively to these targets. In the experiments presented here, N=3 for sake of simplicity. The choice of the articulatory targets is made by activating successively and randomly 3 neurons of the articulatory map. Their preferred vectors code for the articulatory configuration of the target. Finally, gaussian noise is introduced just before sending the target values to the control system. By default, the variance of the gaussian equals 5 percent of the extent of each articulatory dimension.

When an articulation is performed, a model of the vocal tract is used to compute the corresponding acoustic trajectory. There are two models. The first one is abstract and serves as a test model to see which properties are due to the particular shape of the articulatory/acoustic mapping and which are not. This is simply a random linear mapping between he articulatory space and the acoustic space.

The second model is realistic in the sense that it reproduces the human articulatory to perceptual mapping concerning the production of vowels. We model only vowels here for sake of computational efficiency. The three major vowel articulatory parameters are used : (Ladefoged and Maddieson, 1996) tongue height, tongue position and lip rounding. To produce the acoustic output of an articulatory configuration, a simple model of the vocal tract was used, as described in (de Boer, 2000), which generates the first and second effective formants which are

known to represent well human perception of vowels (de Boer, 2000). This model does not allow to deal with consonants, but is enough to investigate at least the phonemic coding of vowel targets.

The experiment presented consists in having a population of agents (typically 20 agents) who are going to interact through the production and perception of sounds. They are endowed with the neural system and one of the articulatory synthesizers described previously. They interact by pairs of two : at each round, one agent is chosen randomly and produces a dynamic articulation according to its articulatory neural map as described earlier. This produces a sound. Then another random agent is chosen, perceives the sound, and updates its neural map with the learning rule described earlier.

Let us describe first what we obtain when agents use the abstract articulator. Initially, as the receptive fields of neurons are randomly and uniformly distributed across the space, the different targets that compose the productions of agents are also randomly and uniformly distributed. What is very interesting, is that this initial state situation is not stable : rapidly, agents get in a situation like on figures 2 which corresponds to figures 1 after 1000 interactions in a population of 20 agents. These shows that the distribution of receptive fields is not anymore uniform but clustered. The associated point attractors are now several, very well-defined, and non-trivial. Moreover, the receptive fields distribution and attractors are approximately the same for all agents. This means that now the targets that agents use belong to one of well-defined clusters, and moreover can be classified automatically as such by the relaxation of the network. In brief, agents produce phonemically coded sounds. The code is the same for all agents at the end of a simulation, but different across simulations due to the inherent stochasticity of the process.

Now, (Oudeyer 2002) showed that when you use the realistic articulatory synthesizer, you get additionally vowel systems (defined as the set of point attractors) which do follow very well the tendencies observed in human languages. As a consequence, this model proposes and show the plausibility of the hypothesis : phonemic coding and the existence of shared categorical systems might be a result of the dynamic properties of very generic neural tissues (the same maps can be used for hand-eye coordination for instance), but which particular categories appear is due to the particular shape of the articulatory to perceptual mapping (but this alone is not necessary for phonemic coding, and we will argue here that is it also not sufficient).

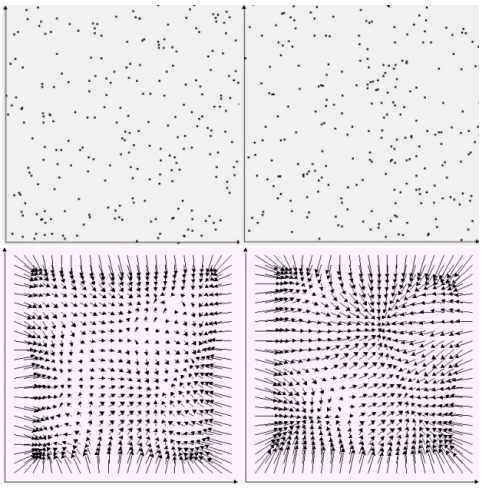

Figure 1: Acoustic 2-D neural maps at the beginning (top), and associated population vector function (bottom) for two agents

Integration within a functional model of the origins of syllable systems

In the coupled neural maps model, there was no shared repertoire of complex sound categorization which was constructed. Obviously, when one agent would hear a complex utterance, the perceptual trajectory would go through zones of the space each belonging to the basin of attraction of a category. But this does not allow to decode appropriately for instance complex utterances made of articulatory targets whose basins of attractions are not connex : the interpolation taking place during the actual articulation will lead to articulatory and perceptual trajectory who go through basins of attraction of categories which do not correspond to any of the initial targets. As a consequence, if one wants to have a model of the origins of syllable systems, it is necessary to add another mechanism. Here this mechanism will be just the one described in (Oudeyer 2001b), and as opposed to the coupled neural maps, is functional.

Basically, agents are going to play the imitation game (de Boer 2000). They possess the two neural maps presented earlier, which work in the same manner. Additionally, each of them have repertoires of syllables (here sequences of N=3 targets), and one game consists in having one agent, the speaker, choose one of its items, then utter it, and then have another random agent, the hearer, try to imitate it by producing the closest syllable in its repertoire. After the imitation, the speaker categorizes the utterance he heard and checks if it corresponds to the category of the syllable he pronounced initially. He then gives a binary (good or bad) feedback to the hearer. The items of their memories have scores (num. of times used successfully / num. of times

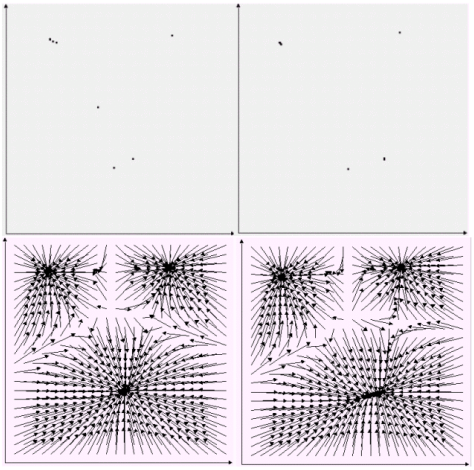

Figure 2: Neural maps and attractors after 1000 interactions, corresponding to the intial states of figure 1)

Figure 3:

used). This is used to prune the syllables which are not efficient. Initially, repertoires are empty. They grow either by imitation (agents hear a syllable that they can not imitate and yet have used a usually efficient syllable prototype), or by invention. Inventing a syllable consist in choosing randomly N=3 targets by activating three random neurons of the articulatory map. As the receptive field of these neurons are initially uniformly spread across the space, inventions at the beginning of simulations will produce syllables whose targets are uniformly spread across the space. This means exactly non phonemically coded.

This model can be tuned to show that only our non-functional hypothesis is plausible in a realistic explanatory setting : this can be done by using a realistic synthesizer or not, coupled with the ability to deactivate the learning rule of neural maps (neurons do not update their receptive fields when they perceive a stimulus). If the use of a realistic synthesizer with a deactivated learning rule does not provide phonemic coding, this does show that the

Figure 4:

morpho-innatist hypothesis is not sufficient, and as (Oudeyer 2002) showed it was not necessary either, the conclusion is that this is not a good candidate (in fact, here we would show even more : the combination of morpho-perceptual constraints and functionalism is not sufficient). If we use an abstract synthesizer with a deactivated learning rule, then if we do not get phonemic coding, Lindlom's functionalism is not sufficient either (and again, we showed it was not necessary). Then, if each time we activate the learning rule we do get phonemic coding, this will confirm the plausibility of the exaptationist hypothesis of (Oudeyer 2002).

A measure of phonemic coding of a syllable systems was developed in (Oudeyer 2002). This consists in making models of the distributions of targets, based on parzen windows : at points corresponding to a the crossings of a regular grid, one approximates the local probability density function by averaging the values of a gaussian function (centered on this point) taken at all points coded by each targets. This is very similar to making multi-dimensional histograms, and counting how many targets fall in each bin. Yet, using gaussians gives a fuzzy binning whose choice of variance is much easier than the choice of bin size in the case of histograms. This approximation is used to compute the entropy of target distributions : if they are uniformly spread, entropy is maximal, and the more they are clustered, the lower the entropy. Thus, this is a way to automatically monitor the clusteredness of targets, and so how much a system is phonemically coded.

A first series of experiment [1] was conducted with the abstract articulatory synthesizer. A first parameter that could be varied (apart from the use or not of the learning rule), was the dimensional-

[1] let us mention that the mechanism described here does allow the cultural building of a syllable systems with which agents can imitate each other successfully. Results identical to what was described in (Oudeyer 2001) were found here. This is a significant progress compared to other other models of the origins of sound systems, but we will not give details here since it is not the main topic this paper

ity of the motor and perceptual spaces. Indeed, this is interesting since the goal of agents is to position syllable prototypes in these spaces such that they are not confused. Hence, the fact that increasing linearly the number of dimensions does increase exponentially the volume of the spaces might have some consequences. Figure 3 presents the average entropies of syllables systems composed of 40 syllables, and generated by populations of agents using or not the learning rule. For each case, 50 experiments were ran and the mean entropy and standard deviation were measures. We added a "uniform" column corresponding to the entropy of syllable systems target distributions generated randomly (uniformly distributed). The values of "uniform" thus characterize syllable systems which are absolutely not phonemically coded. We do observe that except for dimension 1 (which is not very realistic), the use of the learning rule does generate phonemically coded systems while when it is not used, systems have entropies equal to the uniform case : they are not phonemically coded. A second set of experiments consisted in keeping the dimensionality of motor and perceptual spaces equal to 2 and see whether the amount of noise would change anything (noise proved to be of importance in the experiments of de Boer 2000). Figure 4 shows that even a large amount of noise (30 percent) does not push the system to be phonemically coded when it does not use the learning rule, while using it still gives phonemically coded systems when noise gets high (yet it becomes less and less phonemically coded, which is normal since when the noise is to high, this is equivalent to reshuffling permanently and completely all targets).

A second series of similar experiments were made using the realistic articulatory synthesizer. Figure 5 shows the results, when noise is again varied : we see that even if a realistic synthesize is used, no phonemic coding is obtained if the learning rule is not used. And again, as soon as it is used, and if the noise remains reasonable, we do get phonemic coding.

As a consequence, neither morpho-perceptual innatism nor functionalism is sufficient to explain phonemic coding.

Conclusion

(Oudeyer 2002) presented a new hypothesis for the origins of phonemic coding, which had the a priori advantage of being more simple and requiring less assumptions than other models. In this paper we presented a general model of the origins of syllable systems, which in addition to solve the 2-level problems faced by previous research in the origins of sound systems, allows to test all hypothesis on a common ground. We showed that clearly the morpho-perceptual innatism and Lindblom's functionalism are neither necessary nor sufficient to explain phonemic coding. We did not prove that the non-functional side effects of our coupled neural maps model are necessary (but a better hypothesis has to be invented and validated to prove they are not), but they are sufficient, and so the only existing satisfying candidate to explain phonemic coding.

Figure 5:

References

de Boer, B. (2000) The origins of vowel systems, Oxford Linguistics, Oxford University Press.

Chomsky, N. and M. Halle (1968) The Sound Pattern of English. Harper Row, New york.

R. I. Damper (2000) Ontogenetic versus phylogenetic learning in the emergence of phonetic categories. 3rd International Workshop on the Evolution of Language, Paris, France. p.55-58.

Georgopoulos, Kettner, Schwartz (1988), Primate motor cortex and free arm movement to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. Journal of Neurosciences, 8, pp. 2928-2937.

Hurford, J., Studdert-Kennedy M., Knight C. (1998), Approaches to the evolution of language, Cambridge, Cambridge University Press.

Kirby, S. (1998), Syntax without natural selection: how compositionnality emerges from vocabulary in a population of learners, in Hurford, J., Studdert-Kennedy M., Knight C. (eds.), Approaches to the evolution of language, Cambridge, Cambridge University Press.

Kuhl (2000) Language, mind and brain : experience alters perception, The New Cognitive Neurosciences, M. Gazzaniga (ed.), The MIT Press.

Ladefoged, P. and I. Maddison (1996) The Sounds of the World's Languages. Blackwell Publishers, Oxford.

Lindblom, B. (1992) Phonological Units as Adaptive Emergents of Lexical Development, in Ferguson, Menn, Stoel-Gammon (eds.) Phonological Development: Models, Research, Implications, York Press, Timonnium, MD, pp. 565-604.

MacNeilage, P.F. (1998) The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499-548.

Morasso P., Sanguinetti V., Frisone F., Perico L., (1998) Coordinate-free sensorimotor processing: computing with population codes, Neural Networks 11, 1417-1428.

Oudeyer, P-Y. (2001a) Coupled Neural Maps for the Origins of Vowel Systems. Proceedings of ICANN 2001, International Conference on Artificial Neural Networks, Vienna, Austria, LNCS, springer verlag, Lectures Notes in Computer Science, 2001. Springer Verlag.

Oudeyer P-Y. (2001b) The Origins Of Syllable Systems : an Operational Model. in proceedings of the International Conference on Cognitive science, COGSCI'2001, Edinburgh, Scotland., 2001.

Oudeyer, P-Y. (2002) Phonemic coding might be the result of non-functional sensory-motor coupling dynamics, under review in SAB'02.

Stevens, K.N. (1972) The quantal nature of speech : evidence from articulatory-acoustic data, in David, Denes (eds.), Human Communication : a unified view, pp. 51-66, New-York:McGraw-Hill.