Conditions for Stable Vowel Systems in a Population

Bart de Boer

AI-lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium
bartb@arti.vub.ac.be

Abstract. This paper describes an investigation of two computer models of how vowel systems can be transferred from one generation to the next. Humans tend to reduce the articulation of the vowels (and other speech sounds) they produce. If infants would learn on the basis of these reduced signals, vowel systems would collapse rapidly over time. As this is not observed in practice, some mechanism must be present to counter it. Two candidate mechanisms are investigated in this paper: compensatory expansion of articulations learned on the basis of reduced speech sounds and learning on the basis of more carefully articulated speech. It turns out that larger vowel systems with central vowels can only remain stable when learning is based on carefully articulated speech.

1 Introduction

This paper investigates under what circumstances vowel systems are stable when they are transferred from generation to generation in a population of language users. It is obvious that stable transfer occurs. This can be deduced from the observation that children learn to reproduce the vowel systems of their parents as closely as possible in many ways. Hence, for example, the subtle distinctions in pronunciation between closely related dialects, which can be identified readily by speakers. Such stable transfer is not directly obvious. After all, in rapid, casual speech (the kind people use most often) articulation of speech sounds is reduced. This means that the acoustic distance between the different speech sounds becomes less. In vowels this is especially noticeable. The simplest assumption one can make about learning of speech by infants is that they make a statistical analysis of the speech sounds they hear. However, in that case one would expect systems of speech sounds to become slightly more reduced in every new generation, as the most frequently occurring speech is reduced. This would lead to a relatively rapid collapse of the system of speech sounds.

Such collapses are not generally observed in the way languages change over time. Sound systems of languages sometimes change rather rapidly over time (e. g. [8]), but such change is different in nature from the collapse due to reduction in articulation. In the case of the collapse of a vowel system, for example, one would expect vowels to move more and more to a central articulation (such as the vowel in the English word *the*) and different vowels to merge when they get too close together acoustically. In reality one observes complex shifts, mergers and splits of vowels (e.g. [4]). However, these changes preserve the general positions of- and distances between the different vowels in the vowel system.

Therefore the assumption of purely statistical learning must be fallacious. There must be a mechanism that prevents collapse of speech sounds when infants learn them. Two possible mechanisms will be compared in this paper. In order to keep the simulation tractable, only transfer of vowel systems was studied. An extra advantage of vowel systems is that their acquisition and transfer has been relatively well studied.

Both learning mechanisms use the statistical distribution of vowels in the input to determine the position of the vowels that are learned. The first mechanism (the *compensation mechanism*) uses all the available input, but uses the knowledge that the input consists of a reduced version of the original vowel system. It undoes the contraction of the vowel system by applying an expansion that approximates the inverse of the contraction.

The second mechanism (the *infant-directed speech mechanism*) assumes that the input the infant really uses to learn is not as reduced as rapid casual speech. Instead of using all available data for a statistical analysis, the infant only learns on the basis of speech that is articulated more carefully than rapid, casual speech. Such input can be recognized from characteristics of the signal itself (slower speed, clearer intonation, higher volume) or from the setting in which the sound is perceived (one-to-one interactions between the infant and the caretaker, for example).

In a sense, these mechanisms are minimal implementations of learning bias. Bias can be caused by a tendency of a learning mechanism to find representations that are different from the statistical distribution of the input data. It can also be caused by the selection of input data, such that the distribution of the data on which learning is based is different from the distribution of the complete data set. The first possibility is represented by the compensation mechanism; while the second possibility is represented by selective attention to better articulated input data. The algorithms presented below are minimal. The compensation mechanism is the *exact* inverse (except for noise) of the reduction while the selection mechanism only pays attention to the *best* articulations. Any other implementation of these mechanisms must perform less well.

There is *a priori* evidence that both processes could play a role. In order for infants to imitate adults, they must be able to match the sounds they produce with the sounds they perceive from adults. The same vowel produced by an infant and by an adult is quite different acoustically, due to the infant's shorter vocal tract. Therefore infants have to do a normalization of the signals they perceive. If infants are able to do this, it is not unlikely that they are able to do a similar normalization of reduced speech sounds to more expanded versions of the same speech sounds.

Also, there is ample evidence that the kind of speech that is addressed to infants is more carefully articulated than ordinary, adult-directed speech [7]. Such speech is distinguished by slower speed, exaggerated intonation and higher volume. It has been found that infants tend to prefer such speech to ordinary speech [2]. It can therefore be inferred that the kind of input that children use to base their vowel systems on does not necessarily consist of the rapid, casual speech that is used most often between adults. Combined with the fact that this special infant-directed register is found almost universally in different cultures, it would appear that such input plays an important role in the way speech sound systems are learned by children.

This paper uses a computer model to investigate the dynamics of both mechanisms. The computer model is inspired in part by the work on language games [13,14] and in part by the work on the iterated learning model [5,6]. The computer model

consists of a population of agents. All of these agents can produce and perceive speech sounds in a human-like way. Some of these agents model infants and others model adults. Adults talk to infants in a more or less reduced register, and infants learn the adults' vowel system on the basis of the distribution of the input signals they receive. The imperfectness of articulation is modeled by adding noise to the articulatory positions that agents try to achieve. After a while, infants change into adults, old adults are removed from the population and new infants are added to it. It can be monitored over time how the vowel systems in the population change.

2 The simulation

The simulation is based on a population of agents that can interact with each other. Agents exist in two modes: adult mode and infant mode. Adult agents have a fixed repertoire of vowels, while infant agents are busy acquiring a repertoire. Interactions consist of an adult agent producing a vowel from its repertoire, and an infant agent listening to that sound and updating its repertoire of speech sounds. Each agent participating in the interaction is chosen at random. After a certain number of interactions, the adult agents are removed from the population, infant agents become adult agents and a new batch of infant agents is added to the population. The agent's vowel systems are logged and it can be investigated how they change over time.

The agents can produce and perceive speech sounds in a human-like way. For this purpose they are equipped with a vowel synthesizer and a means to calculate distances between vowels. The vowel synthesizer is the same as the one used in [1]. It has three articulatory inputs: the position of the tongue in the front-back dimension, the height of the tongue and the amount of rounding of the lips. These correspond to the parameters that phoneticians usually employ to describe vowels [9, ch. 9]. In the model each input can have values between 0 and 1, where 0 corresponds to most front, lowest and least rounded, while 1 correspond to most back, highest and most rounded. Thus the vowel [a] can be described by the parameter values (0, 0, 0) while the vowel [u] can be described by the parameter values (1, 1, 1). Its outputs consist of the frequencies of the first four formants (resonances of the vocal tract). For example, for the inputs (0, 0, 0), the output would be (708, 1517, 2427, 3678) and for the inputs (1, 1, 1) the output would be (276, 740, 2177, 3506). With this articulatory model all ordinary vowels can be generated

The perception function is very similar to the one used in [1] but it follows the original perception model by Schwartz *et al*. [12] more closely. It is based on the observation that most vowel signals can be simplified with what is called an effective second formant. Vowel signals generally have multiple peaks in their frequency spectrum, each peak corresponding with a resonance frequency of the vocal tract. However, one can generate an artificial signal that sounds very similar to the original vowel using a frequency spectrum that has only two peaks. The position of the first peak in such a spectrum is equal to the position of the first peak in the original spectrum, but the position of the second peak is a non-linearly weighted sum of the positions of the second, third and fourth peaks in the original spectrum. This second peak is called the *effective second formant*.

The position of the effective second formant can be calculated using a formula due to Mantakas *et al.* [11] that was adapted by [12]. In the perception model, frequencies in Hertz are converted to the more perceptually inspired Bark frequency scale. Equal distances in the Bark scale correspond to equal perceived differences in pitch.

The conversion is performed with the following formula:

$$Bark = 7 \cdot \sinh^{-1}\left(Hertz/650\right)$$

Given the first formant and the effective second formant in Barks, the perceptual distance between two vowels *a* and *b* can then be calculated as follows:

$$D = \sqrt{\left(F_{1,a} - F_{1,b}\right)^2 + \lambda^2 \left(F'_{2,a} - F'_{2,b}\right)^2}$$

where F_1 is the first formant and F'_2 the effective second formant (of vowels *a* and *b*, respectively). *D* is the perceptual distance and λ is a constant that regulates the importance of the effective second formant relative to the first formant. It has a value of 0.3 in all experiments presented here, a value which has been found to generate perceptually realistic distances [12].

Agents engage in interactions. For each interaction an adult agent and an infant agent are chosen randomly from the population. In all experiments presented here, the population contains 20 adult agents and 20 infant agents. An adult chooses a random vowel from its repertoire and produces this. The infant perceives this signal and uses it to derive the possible vowels that are used in the population. In all the experiments, 10 000 interactions are performed, after which all the adult agents are removed from the population, and the infant agents turn into adults.

As has been mentioned above, adults can produce utterances from their repertoire of vowels. For each of the agent's vowels the articulatory parameters are stored. However, humans cannot produce utterances perfectly. In order to model this, noise is added to the articulatory parameters before they are synthesised. For each of the articulatory parameters, this noise is taken from the normal distribution with mean 0 and standard deviation 0.05. Also, the sloppiness of agents' articulations can be modelled by adding a bias for articulations to become more centralized. This is done in the following way:

$$x_{sloppy} \leftarrow x + \alpha\left(0.5 - x\right) \tag{1}$$

where *x* is any of the articulators (position, height or rounding) and α is a constant that determines how much the articulations are attracted towards the centre. If α is 0, no reduction takes place, and if α is 1 all articulations are reduced to 0.5. A realistic value for α is relatively hard to determine for rapid, casual speech, as it is not known where the original articulations are supposed to lie, but a realistic value would be at least 20% (0.2). This can be deduced from the fact that the surface of the acoustic space (F_1-F'_2 space) that is used for infant-directed speech, which can be considered carefully articulated, is at least 1.4 times the surface that is used for ordinary adult-directed speech (Hiu Mei Liu, *personal communication*). This amounts to a linear reduction of about 20%.

While the adult agents produce sounds, the infant agents perceive and learn speech sounds. In reality, learning must occur incrementally in infants. This means that each sound an infant perceives must slightly update the representations in its brain until the

different vowel categories are represented. This could in principle be modelled with a neural network, but for simplicity of implementation, it was decided to use a batch-learning algorithm. Thus infant agents participate in a number of interactions and store all the signals that were used in those interactions. Then, when they are converted into adults, they derive the vowel categories from these stored signals.

Vowel categories are derived with an unsupervised classification algorithm called iterative valley seeking [3]. Iterative valley seeking is a parameter-free classification algorithm. This means that it makes no assumptions about the distribution that underlies the observed data points. It assumes that each peak in a distribution corresponds with a category, and it tries to locate these peaks by using the local density of data points. In order to estimate the local density, the method uses one parameter: the radius R of each point's local neighbourhood. This parameter determines how much the distribution of data points is smoothed when determining the peaks.

This process works well in practice, and is not extremely sensitive to the value of R. A value of 0.3 was used here. However, in the simulations there existed a tendency for outliers to be interpreted as real vowels. These vowels were then produced when the infant became an adult, and spread rapidly through the population. In order to prevent such spurious vowels to spread through the population, a requirement was added that classes needed to contain a minimum number of data points before they were accepted as real vowels. In order to determine whether to accept a class as representing a genuine vowel or not, first the average number of data points per class was calculated. Every class that had more than one third of the average number of data points was accepted as a genuine vowel.

Iterative valley seeking and the selection criterion result in a number of classes with example data points, but not yet in an acoustic or articulatory representation of a vowel. The acoustic representation was determined directly from the data points of the class. It was decided that the best way to determine the acoustic prototype of a class was to take the point at which the density of the class was highest. As no assumptions could be made about the distribution of the data points, the K-nearest neighbour method was used, with $K = 3$. This means that the densest part of the class was assumed to lie at the point that had the nearest third neighbour. This simple method turns out to work well for the kinds of distributions and the number of data points used here.

The articulatory prototype corresponding to this acoustic data point was then found by the agent talking to itself. It started with a vowel at articulatory position (0.5, 0.5, 0.5). Small modifications were made to this, these were articulated and it was measured whether they moved the acoustic signal closer to the target signal. This process was iterated until no more improvement could be achieved. The size for each modification step was a random value from the normal distribution with mean 0.1 and standard deviation 0.01 .

In the implementation of the compensation mechanism, agents re-expand their vowel repertoire in order to compensate for the reduced articulation of the adult agents. This expansion is the opposite of the reduction described in equation (1):

$$x_{expanded} \leftarrow x - \beta(0.5 - x)$$

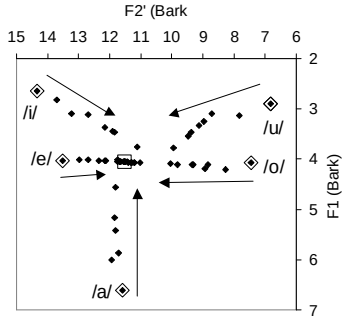

Fig. 1 Collapse of a five-vowel system under 20% reduction

where β is a constant that should have the value $\alpha/(1-\alpha)$ in order to compensate exactly for a reduction of size α. In all experiments presented here that use expansion, the value of β is chosen to compensate exactly for the value of α.

These methods result in reliable learning of a vowel system by a population of infant agents. However, the intention of the research was to check whether vowel systems would be stable when transfer was repeated over many generations.

3 Experiments

In the experiments, the stability of vowel systems over time was investigated. First, it had to be established that vowel systems do indeed collapse when learning is based on reduced speech. Therefore, the evolution over time of a five-vowel system (consisting of /i/, /e/, /a/, /o/ and /u/) was investigated when a minimally realistic amount of 20% reduction was present. The first generation of adults was initialized with the original vowel system, and this was transferred to successive populations of infants. As can be seen in figure 1 the system collapsed to a one-vowel system within 25 generations. In this figure, the vowel system of one agent from the population is plotted for each generation. The first- and effective second formant are plotted in such a way that the position of the vowels correspond to the way phoneticians usually plot them. The open carets represent the original vowels, the black dots vowels in subsequent generations and the open square the final vowel. This result establishes the necessity of an alternative mechanism to preserve vowel systems.

The first experiments were done with the same five-vowel system. The compensation mechanism was implemented with the same 20% reduction, but now with a 25% expansion that, in the ideal case, would exactly compensate the reduction. The infant-directed speech mechanism was implemented as a reduction of only 2%. Some reduction was used, as it cannot be assumed that infant-directed speech is perfect. However, no expansion was assumed. It turned out that no significant difference could be observed between the two mechanisms and that vowel systems were almost perfectly preserved over runs of 100 generations. An example of vowel system preservation in

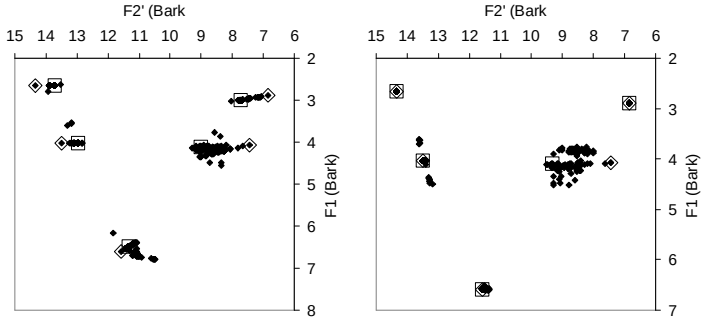

Fig. 2 Evolution of five vowel systems over time. The left frame shows the effect of the compensation mechanism and the right frame shows the effect of the infant-directed speech mechanism

both cases is presented in figure 2. Again starting positions are indicated with open carets, intermediate ones with black dots and final ones with open squares.

The case of seven-vowel systems was more interesting. Three different seven-vowel systems were investigated. All these systems contained the vowels /i/, /e/, /a/, /o/ and /u/. One system (type 1) contained /ɛ/ (as in “bed”) and /ɔ/ (as in “pot”). Type 2 contained /y/ and /ø/ (front rounded vowels, as occur in French end German), while type 3 contained the vowels /ɯ/ (Russian “ы” or Turkish “ı”) and /ə/ (the vowel in English “the”).

In order to obtain a statistically significant comparison of the two mechanisms when transferring these vowel systems, the number of vowels per agent for each generation was monitored. This turned out to be a good measure of vowel system preservation. The evolution of the different types of vowel systems over 250 generations is shown in figure 3. Black lines indicate performance of the infant-directed speech mechanism while gray (orange) lines indicate the performance of the compensation mechanism. Neither of these mechanisms is better than the other in all cases. Also, for both mechanisms, significant reduction of the vowel systems is observed. Apparently, neither of the mechanisms is sufficient for preserving seven-vowel systems.

The combination of both mechanisms was therefore tested. This was implemented

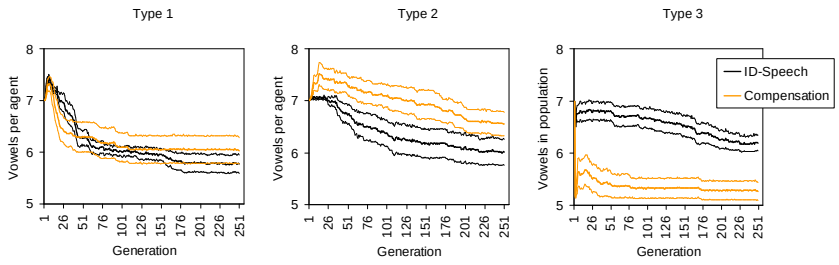

Fig. 3 Change of seven-vowel systems over time. Thick lines indicate averages, thin lines indicate 95% confidence intervals

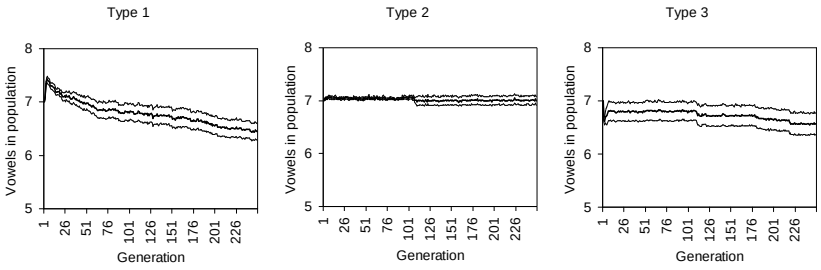

Fig. 4 Evolution of average vowel system size when both compensation and infant-directed speech are used. Dotted lines indicate 95% confidence intervals

by using a reduction of 2% and a corresponding expansion of $100/49 \approx 2.04\%$. As can be observed in figure 4 this resulted in much better preservation of vowel systems over time. Although vowel systems are still reduced over time, this happens so slowly that it could be realistic. It should also be kept in mind that the learning mechanism was constructed to avoid learning vowels that occurred infrequently. Once a vowel is lost, it is therefore not possible to relearn it, so that the system is biased towards shrinking vowel systems over time.

4 Conclusion and Discussion

Infants learn vowel systems on the basis of the sounds that their parents produce. However, the most frequent register of adult speech, rapid casual speech, is often very much reduced. It is therefore clear that infants cannot learn the categories of speech sounds on the basis of an unbiased statistical analysis of the speech signals they perceive. This would result in a rapid collapse of the language's vowel system, as was shown in the first experiments presented above.

In the other experiments, two minimally sufficient models of preserving structure in vowel systems were compared. Although the exact behaviour of the models is of course dependent on the details of the learning algortihms, they were kept as simple as possible, and in a sense "minimal"; any other learning method would give less good learning results. Also, the simulation study was a qualitative study, and only measured the stability of vowel systems over time. No attempt was made to model the exact duration and trajectory of collapse of real human vowel systems. Therefore it is probably safe to draw qualitative conclusions on the basis of the models.

Both methods were based on statistical learning of the input data. In the first method, infant agents learned the number and the position of vowel categories on the basis of speech that was reduced by about 20%. However, the infant agents compensated for this by pushing the learned vowel categories away from the center of the articulatory space. In the second method, infant agents learned the position of vowel categories on the basis of speech that was only slightly reduced (2%). The infant agents did not compensate for the reduced articulation.

In the absence of experimental evidence, the method by which infants compensate is to be preferred, as infants have to perform renormalization of speech sounds in any case. The infant vocal tract produces signals that are quite different from those produced by the adult vocal tract for similar articulatory gestures. In order to learn to imitate speech sounds, infants need to be able to compensate for this. If such compensation needs to take place, an extra compensation to account for reduced articulation can be assumed to take place as well.

The criterion for comparison of the different methods is the number of generations over which vowel systems are preserved in the population. Although vowel systems of human languages can change rather rapidly [8] complex vowel systems of languages can remain stable for longer periods of time as well. Such stability must be present before other historic processes (influence of phonetic context, for example) can change vowel systems. A reasonable threshold for stability is preservation of a vowel system over approximately 50 generations, which corresponds to 800–1000 years in a (prehistoric) human population. After such a period of time, different historic processes generally have changed human vowel systems.

It was shown in the experiments that both methods were perfectly able to preserve five-vowel systems without central vowels. Such systems remained stable over at least 100 generations. On the basis of this data, no conclusion can therefore be drawn about which method best explains learning by infant. When learning of seven-vowel systems was tested, it was found that neither mechanism alone was able to assure stable transfer in all cases. However, a combination of both mechanisms was able to achieve stable transfer of vowel systems.

Either mechanism can account for learning of small vowel systems. As has been outlined above, the compensation mechanism is to be preferred in such a case. Although a more sophisticated compensation mechanism could probably be designed to preserve stability of more complex vowel systems, each automatic compensation mechanism must have vowel combinations that it cannot successfully reconstruct from the reduced input. After all, information does get lost in the reduction. Possibly compensatory mechanisms for seven vowel systems can still be designed, but human vowel systems can have up to at least 17 different vowels (Norwegian [15]), and given the experimental results, it is unlikely that such complex systems can be stably reconstructed from reduced articulations.

Therefore, for successful transmission of complex vowel systems, learning needs to be based on more carefully articulated examples as well. This would indicate that special infant-directed speech registers (motherese) are more important when a language has more vowels. Whether this really is the case has not been experimentally investigated so far, but data from a study of infant-directed speech in different languages [7] seems to indicate that the more vowels a language has, the more carefully articulated special infant-directed speech is. From the data in [7], it can be measured how much more area of the 2-D acoustic space is used for infant-directed (ID) speech than for adult-directed (AD) speech. It turns out that Russian, with a small (six-) vowel system has an ID/AD ratio of 1.73, English, with an intermediate size vowel system has a ratio of 1.85 and Swedish, with a large vowel system has a ratio of 1.96. Additional data [10] shows that Mandarin Chinese with a vowel system that is slightly smaller than that of Russian has a ratio of about 1.4. These data seem to indicate that languages with large vowel systems use more carefully articulated infant-directed

speech. This data corresponds well with the finding of this paper that vowel systems with larger number of vowels need carefully articulated examples to be transferred from one generation to the next.

The conclusion that can be drawn from the experiments presented here as well as from the infant-directed and adult-directed speech data is that carefully articulated examples are not necessary as long as small and simple vowel systems need to be learned. Compensation for reduced articulation can be performed by simple expansion of the learned vowel prototypes. However, more complex vowel systems can only be learned successfully when more carefully articulated examples are available. Such carefully articulated examples can be found in infant-directed speech, and it is found that this speech register is more pronounced in languages with larger vowel systems.

References

1. de Boer, B.: Self-organization in vowel systems, Journal of Phonetics 28(4), (2000) 441–465
2. Fernald, A.: Four month-old infants prefer to listen to motherese. Infant Behavior and Development 8, (1985) 181–195
3. Fukunaga, K.: Introduction to statistical pattern recognition, Boston: Academic Press (1990).
4. Hock, H. H.: Principles of Historical Linguistics, second edition, Berlin: Mouton de Gruyter. (1991)
5. Kirby, S.: Function, Selection and Innateness: The emergence of language universals, Oxford: Oxford University Press. (1999)
6. Kirby, S.: Natural Language from Artificial Life, Artificial Life 8 (2002) 185–215
7. Kuhl, P. K., Andruski J. E., Chistovich, I. A., Chistovich, L. A. Kozhevikova, E. V., Rysinka, V. L., Stolyarova, E. I., Sundberg, U. & Lacerda, F.: Cross-Language Analysis of Phonetic Units in Language Addressed to Infants, Science 277 (1997) 684–686
8. Labov, W.: Principles of Linguistic Change: internal factors, Oxford:Blackwell (1994)
9. Ladefoged, P. & Maddieson, I.: The Sounds of the World's Languages, Oxford: Blackwell (1996).
10. Liu, H.-M., Tsao, F.-M. & Kuhl, P. K. Support for an expanded vowel triangle in Mandarin motherese. International Journal of Psychology, 35(3–4) (2000) 337
11. Mantakas, M, J.L. Schwartz & P. Escudier Modèle de prédiction du 'deuxiéme formant effectif' F2'—application à l'étude de la labialité des voyelles avant du français. In Proceesings of the 15th journées d'étude sur la parole. Société Française d'Acoustique, (1986) 157–161.
12. Schwartz, J.-L., Boë, L.-J., Vallée, N. & Abry, C.: The Dispersion-Focalization Theory of vowel systems. Journal of Phonetics 25, (1997) 255-286.
13. Steels, L.: A Self-Organizing Spatial Vocabulary. Artificial Life 2(3) (1995) 319–332.
14. Steels, L.: The Synthetic Modelling of Language Origins, Evolution of Communication 1(1) (1997) 1–34.
15. Vanvik, A.: A phonetic-phonemic analysis of Standard Eastern Norwegian. Norwegian Journal of Linguistics 26 (1972) 119–64.