How to keep a vowel system stable over time

Abstract

This paper describes an investigation of two 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 this. 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.

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 has been observed that in rapid, casual speech (the kind people use most often) speech sounds are reduced. This means that they are not articulated completely anymore, and therefore that the acoustic distance between the different speech sounds is reduced. This is especially noticeable in vowels. 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 that can be identified readily by speakers, but that do not impede mutual understanding. However, if children would base the vowel system they learn directly on the basis of the statistical distribution of vowels in their input, one would expect that such systems would become slightly more reduced in every generation, leading to a relatively rapid collapse of a language's vowel system. The same would most likely be true for its consonant system as well.

However, such collapses are not observed in the way languages change over time. Sound systems of languages can change rather rapidly over time (e. g. Labov 1994) in contrast to what was formerly believed, but such change is rather different from the collapse of a vowel system due to reduction in articulation. In such a case, one would expect vowels to move more and more to a central articulation (such the vowel in the English word *the*) and different vowels to merge when they get too close together acoustically. In contrast, in reality one observes complex shifts, mergers and splits of vowels (e.g. Hock 1991). However, these changes preserve the general positions of- and distances between the different vowels in the vowel system.

Therefore there must be a mechanism that prevents collapse of a vowel system when it is learned by infants. Two possible mechanisms will be compared in this paper. Both mechanisms use the statistical distribution of the vowels in the input the infant receives to determine the position of the vowels the infant learns. The first mechanism uses all the available input, but uses the knowledge that the input consists of a reduced version of the original vowel system to expand the vowel system that is ultimately learnt. In a sense it undoes the contraction of the vowel system by applying an expansion that is the inverse of the contraction.

The second mechanism assumes that somehow the input to the infant uses to learn its vowel system is not as reduced as rapid casual speech, but that the available input that contains better examples of the vowels to be learnt than rapid casual speech. Such

input can be recognized from characteristics of the input 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).
A priori there is evidence that both processes could play a role. In order for infants to be able to imitate adults, they must be able to renormalize the sounds they produce and the sounds they perceive from adults. Acoustically, the same vowel produced by an infant and by an adult are quite different, due to the infant's shorter vocal tract. And if infants are able to do such a renormalization, it is not unlikely that they are able to do a renormalization of a reduced speech sound to a more expanded version of the same speech sound.
Also, there is ample evidence that the kind of speech that is addressed to infants, called infant-directed (ID) speech or motherese, is more carefully articulated than ordinary, adult-directed (AD) speech (Kuhl *et al.* 1997). Such speech is distinguished by slower speed, exaggerated intonation and higher volume. It has been found (Fernald 1985) that infants tend to prefer such speech to ordinary speech. It can therefore be inferred that the kind of input that children use to base their vowel systems on does not 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, this would indicate that such input plays an important role in the way speech sound systems are learnt by children.
This paper uses a computer model to investigate the dynamics of both methods. The computer model is inspired in part by the work on language games (Steels 1995, 1997) and in part by the work on the iterated learning model (Kirby 1999; 2002). 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 children in a more or less reduced register, and children learn the adult's 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 followed over time how the vowel systems in the population change.
The intention of this study is not to choose between these two mechanisms. It is quite likely that both mechanisms play a role simultaneously. However, the intention was to determine the ways in which sound systems change over time with both mechanisms and compare this with what is observed in real language, in order to gain more insight into the exact dynamics of the transfer of language from one generation to the next.

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. 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 on the basis of the sounds it heard. 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 (de Boer 2000). 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 (ladefoged & Maddieson 1996, 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 t 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 (de Boer 2000) but it follows the original perception model by Schwartz *et al*. (1997) 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 generated an artificial signal that sounds very similar to the original vowel using a frequency spectrum that only has 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.* (1986) that was later adapted by Schwartz *et al*. (1997). All formulas and algorithms in the perception model are taken from Schwartz *et al*. (1997). In order to make this calculation, the frequencies in Hertz are first 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:

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

The value for the effective second formant frequency can then be calculated with algorithm 1.

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 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 (Schwartz *et al*. 1997).

Algorithm 1: Calculation of the effective second formant. Note that inputs and outputs are in Barks.

$F_2' \leftarrow$ **Calculate effective second formant**(F_2, F_3, F_4)

if $F_2 - F_3 > 3.5$
  $F_2' \leftarrow F_2$
else if $F_3 - F_2 > 2.5$
  $c \leftarrow 3.5 - (F_3 - F_2)$
  $F_2' \leftarrow \frac{F_2 + 0.5cF_3}{1 + 0.5c}$
else if $F_4 - F_2 > 3.5 \vee F_4 - F_3 > F_3 - F_2$
  $F_2' \leftarrow \frac{F_2 + 0.5F_3}{1.5}$
else
  $F_2' \leftarrow \frac{F_3 + 0.5F_4}{1.5}$

Agents engage in interactions. For each interaction one adult agent and one infant agent is chosen randomly from the population. In all the experiments presented here, the population contains 20 adult agents and 20 infant agents. An adult agent chooses a random vowel from its repertoire and produces this. The infant agent perceives this signal and uses it to derive the possible vowels that are used in the language of 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, an agent cannot produce an utterance 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:

$$x_{sloppy} \leftarrow x + \alpha(0.5 - x) \quad (0)$$

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 attraction 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 somewhere over 20%. 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%. However, the value of α will be varied in the different experiments presented below. Figure 2 shows the effect of different values of a on the distribution of signals in the acoustic space.

While the adult agents produce sounds, the infant agents perceive and learn speech sounds. In reality, learning must occur incrementally in infants, that is, each sound they perceive slightly updates the representations in their brains until the different

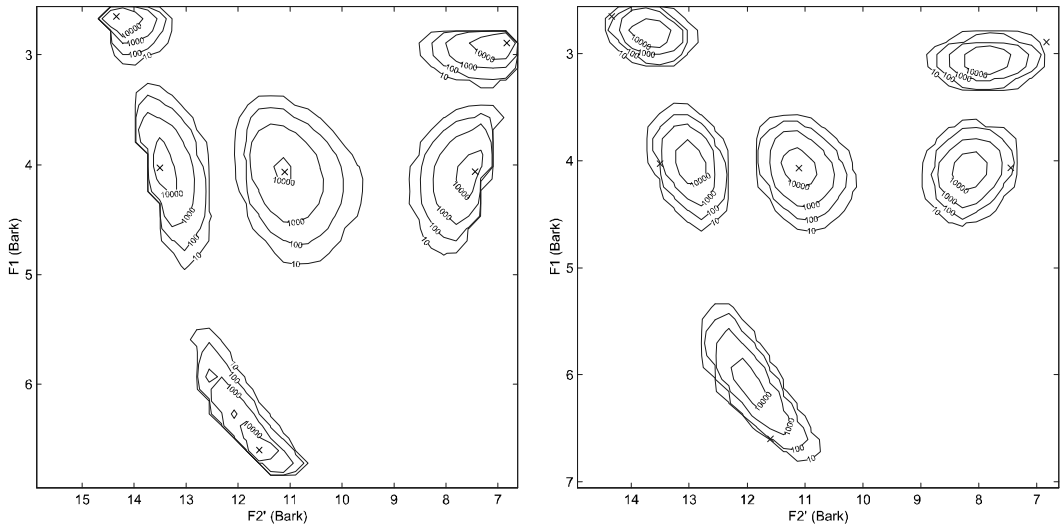

Figure 2: Density of distribution of signals generated by an agent with (left) accurate pronunciation (α = 0.02) and (right) with sloppy pronunciation (α = 0.20). The agents both use the same set of vowels [i], [e], [a], [o], [u] and [ə] which are indicated with crosses. Articulatory noise has standard deviation 0.05. Lines indicate equal densities, and the figures on the lines give an arbitrary measure of the relative density at the line.

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 where 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 (Fukunaga 1990). 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 in an indirect way how much the distribution of data points is smoothed

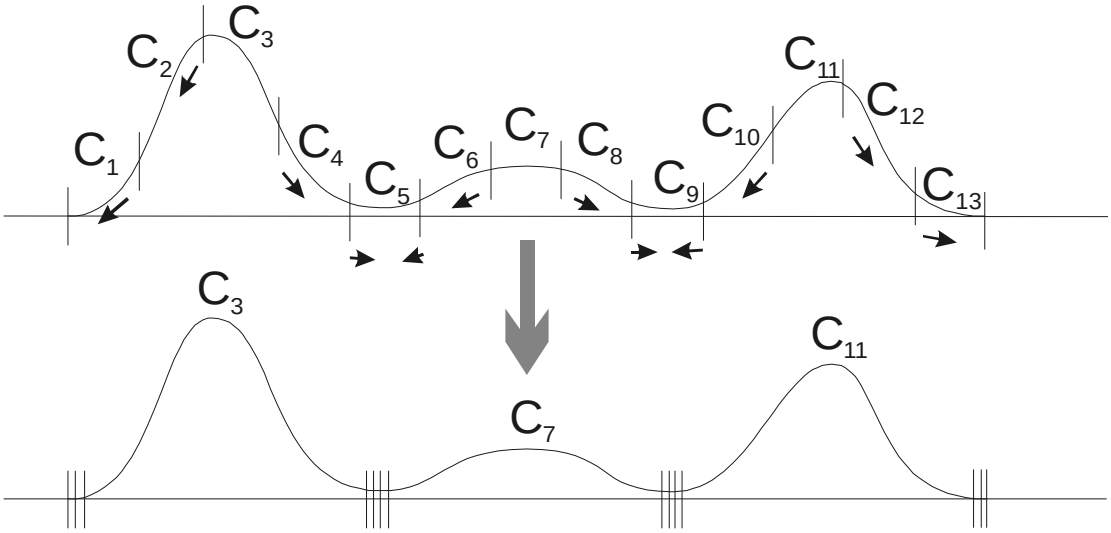

Figure 1: Iterative valley seeking in action. From a "random" initialization, category boundaries are pushed into the valleys of the distribution. Eventually, only the categories at the peaks of the distribution remain large enough to contain points in the sample.

Algorithm 2: Algorithms for learning vowels from perceived signals: overall learning algorithm, iterative valley seeking algorithm and initial classification algorithm.

Learn Phonemes (S)
[S is set of unlabeled acoustic signals]

$C \leftarrow$ **Unsupervised Classify** (S, R)
[C is set of labelled acoustic signals,
R is the radius of the neighbourhoods]

$C \leftarrow$ **Prune Classes**(C)

$P \leftarrow$ **Find Densest Points**(C)
[P is set of acoustic signals]

$V \leftarrow$ **Find Articulations**(P)
[V is set of articulations]

Store(V)

$C \leftarrow$ **Unsupervised Classify** (S, R)

$C \leftarrow$ **Initial Classification**(S, D)
for $\forall c_i \in C$:
$N_i \leftarrow \varnothing$ [N_i is neighbourhood of c_i]
for $\forall c_j \in C$, $c_j \neq c_i$:
if distance(c_i, c_j) $\leq R$
$N_i \leftarrow N_i \cup c_j$
end if
end for
end for

do
for $\forall c_i \in C$ simultaneously**:**
$l_f \leftarrow$ most frequent class label in N_i
set label of c_i to l_f
end for
while(C changed $\wedge$ iterations < 128)

$C \leftarrow$ **Initial Classification**(S, D)
[D is the number of divisions per formant]

for $i \leftarrow 1$ to 4
$maxf(i) \leftarrow$ **maximum of** F_i over $\forall\ s_i \in S$
$minf(i) \leftarrow$ **minimum of** F_i over $\forall\ s_i \in S$
end for

for $\forall s_j \in S$:
acoustic signal of $c_j \leftarrow s_j$
label $c_j \leftarrow 0$
for $i \leftarrow 1$ to 4

$$\text{label } c_j \leftarrow c_j + D\left\lfloor \frac{D-1}{maxf(i) - minf(i)} \left(F_i(s_j) - minf(i)\right) + 0.5 \right\rfloor$$

end for
end for

when determining the peaks.

Iterative valley seeking starts by assigning random class labels to the data points. In the implementation used here, these labels are not really random, but they depend on the position of the data points in the acoustic space. Then for each point it is determined which class label occurs most often within distance R. Finally, the label of each point is set to the label that occurs most frequently in its neighbourhood. This process is iterated. In this way, points tend to get the label of the class that has the highest density in its neighbourhood. In other words, peaks in the distribution tend to become inhabited by only one class, and all the other classes move into the valleys, until these classes become empty of data points. This process is illustrated in figure 1. The process terminates either when the class assignement of data points does not change anymore, or when a maximum number of iterations (128 in all simulations presented here) was reached.

This process works well in practice, does not depend extremely on the value of R. However, in the simulations presented here, there existed a tendency for outliers to be interpreted by infants as real vowels. These vowels would then be generated when the infant became an adult, and so spread rapidly through the population. In order to prevent such spurious phonemes to spread through the population, a requirement was added that classes needed to contain a minimum number of data points before they would be accepted as vowel categories. 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 parameter-free *K*-nearest neighbour method was used, with $K = 3$. This means that the densest part of the class was supposed 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 investigated whether they would move the acoustic signal closer to the target signal. This process

Algorithm 3: Algorithms for finding number and position of classes as well as finding the correct articulatory positions.

$C' \leftarrow$ **Prune Classes**(C)

$a \leftarrow$ mean number of data points per class
$C' \leftarrow \varnothing$
for $\forall c_i \in C$:
  if $\|$members of class of $c_i\| \geq a/3$
    $C' \leftarrow C' \cup c_i$
  end if
end for

$P \leftarrow$ **Find Densest Points**(C)

$P \leftarrow \varnothing$
for all class labels l from C:
  for $\forall c \in C$ with label l:
    Find point c_l that has the closest 3rd-nearest neighbour
  end for
  $P \leftarrow P \cup c_l$
end for

$V \leftarrow$ **Find Articulations**(P)

for $\forall p_i \in P$:
  $v_i \leftarrow (0.5, 0.5, 0.5)$
  do
    $v_i \leftarrow$ **Shift Closer**(v_i, p_i)
  while(improvement)
end for

$v' \leftarrow$**Shift Closer**(v, p)

generate 6 elements of neighbourhood N of v by adding and subtracting a value from distribution $\mathfrak{N}(0.1, 0.01)$ to each articulatory parameter
for $\forall v_n \in N$
  find v_{best} whose signal is closest to p
end for
if signal of v_{best} is better than signal of v
  $v' \leftarrow v_{best}$
else
  $v' \leftarrow v$
end if

was iterated until no more improvement could be achieved. The size for each modification step was a random value from the standard distribution with mean 0.1

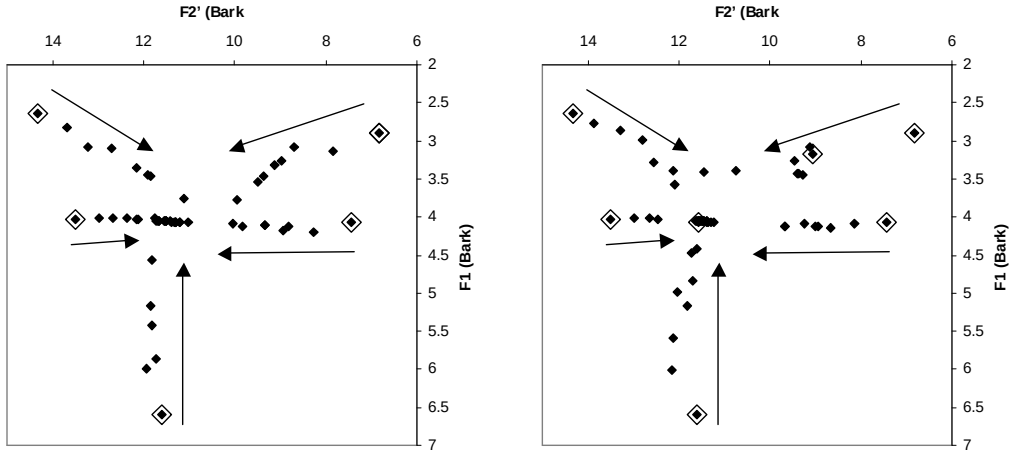

Figure 3: Collapse of vowel systems over 25 generations for a five-vowel (left) and seven-vowel (right) system. Large white lozenges indicate starting positions of vowels. Arrows indicate direction of collapse.

and standard deviation 0.01. All of this is presented in pseudo code in algorithms 2 and 3.

After learning, in one variant of the model, agents can re-expand their vowel repertoire in order to compensate for the reduction that the adult agents have in their articulation. This expansion is the opposite of the reduction described in equation 0:

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

Where β is a constant that should have the value $\frac{\alpha}{1-\alpha}$ 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.

Experiments

The experiments were intended to investigate the difference between a population in which the agents learned on the basis of carefully pronounced examples and a population where agents learned on the basis of less carefully pronounced examples, but where they compensated for the reduced examples by expanding their vowel systems internally. The criterion for quality of transfer was the speed with which the vowel system changed while being transferred over the generations. Both models were investigated with a five-vowel system and a seven-vowel system. The five-vowel system consisted of the vowels [i], [e], [a], [o] and [u]. The seven vowel system had two extra central vowels [ɯ] and [ə].

First of all, it should be established that vowel systems do not remain stable in a population where learning takes place on the basis of reduced pronunciation. In the case of a reduction rate of 0.2, vowel systems in a population collapse within less than 25 generations. This is illustrated in figure 3 and figure 4. Figure 3 shows how the vowel systems of individual agents change over time. For this figure, the vowel system of one agent from each generation was plotted in F1-F2' space. It can be seen that over time, the vowels move towards the center of the acoustic space, and that in

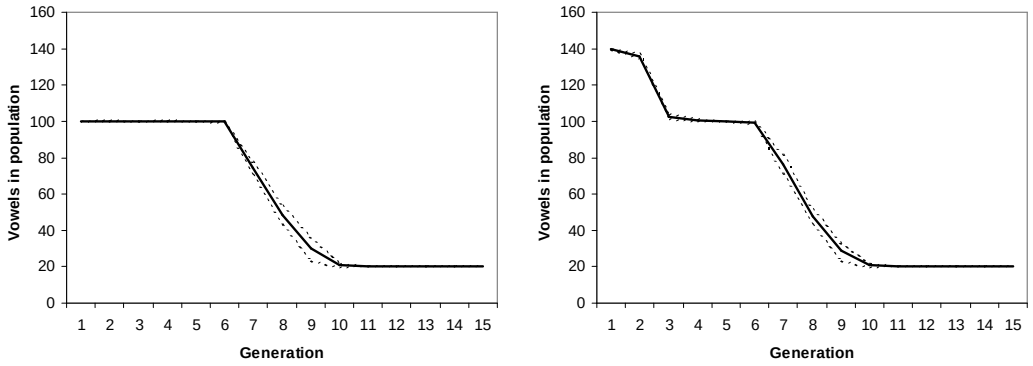

Figure 4: Collapse of the number of vowels in populations with initially five (left) and seven (right) vowels per agent. Each population consists of twenty agents. Shown is the average over ten runs of the system (fat line) as well as the 95% confidence interval (dotted lines).

the end only one central vowel is left over. The figure shows vowel systems over 25 generations for systems starting with five (left) and seven (right) vowels.

Figure 4 shows the evolution over the generations of the average number of vowels in the population, as well as the 95% confidence interval. The average and the confidence interval were calculated over ten different runs of the system. Each of the runs was initialized with the same vowel system, but used different random numbers to calculate articulatory and acoustic noise as well as which agents were picked for interactions. Again the left part of the figure shows the results for a population that initially has a five-vowel system, and the right part shows the results for a seven-vowel system.

From both figures it is clear that vowel systems collapse very quickly (in about ten generations) when learning is based on reduced pronunciation and no compensation is made. This was to be expected.

The first model investigated is the one in which adult agents reduce their pronunciation while infant agents compensate for this. In a sense reduced production shrinks the available (acoustic and articulatory) space for vowels, while the learning mechanism actively expands this space to compensate for the shrinkage. The shrinkage parameter was set to 0.2 and the expansion parameter was set to 0.25, such

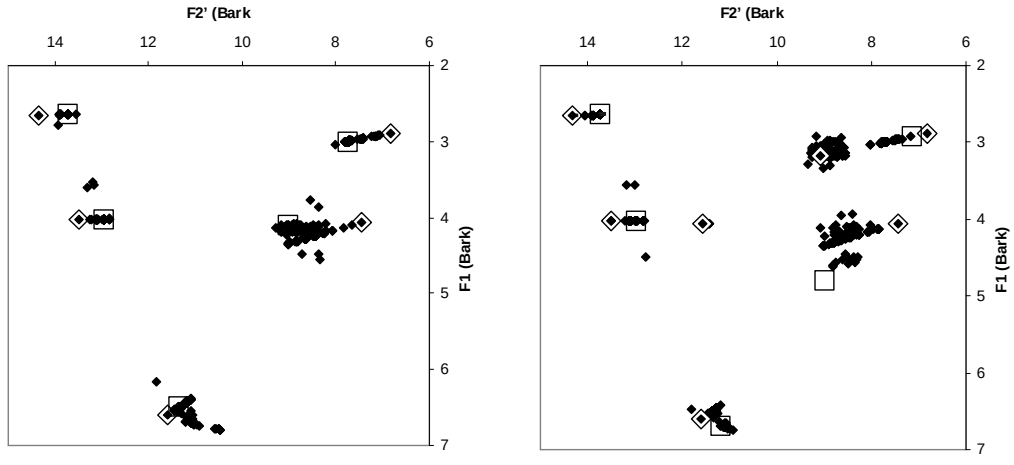

Figure 5: Change over time of vowel systems in a population in which infants compensate for reduced articulation. Shown are vowel systems of individual agents from 100 generations from populations that started with five (left) and seven (right) vowel systems.

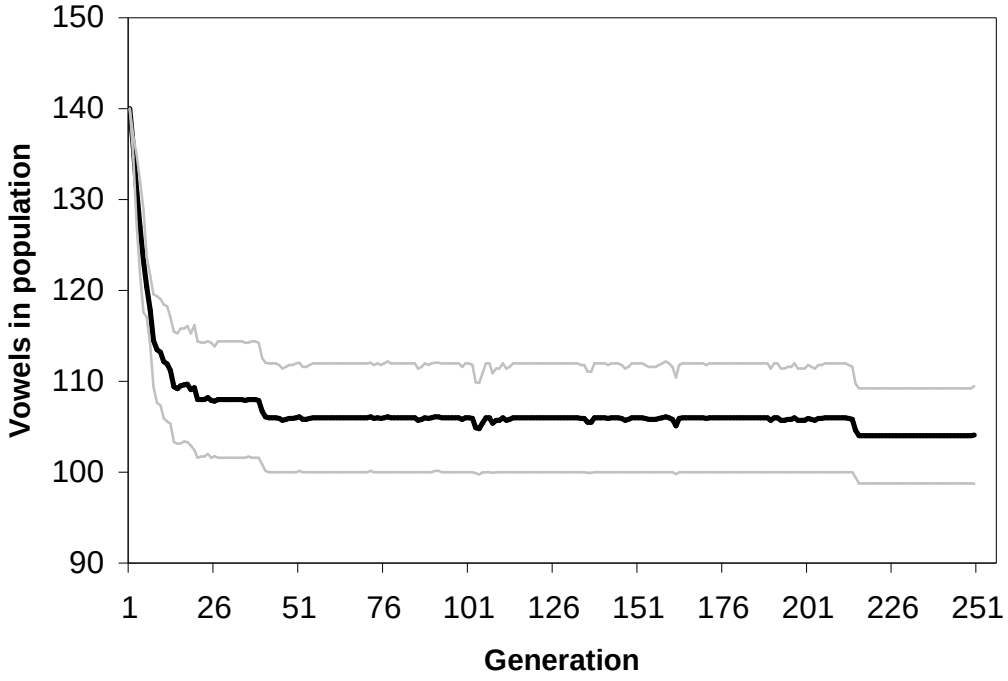

Figure 6: Number of vowels over time and 95% confidence interval for transfer of vowel systems, starting with a seven-vowel system initially.

that (in theory) they compensated exactly for each other. The results of running the computer models a large number of generations are presented in figures 5 and 6. After 100 generations, there were no significant changes to the vowel systems in the cases shown in figure 5. This figure, analogously to figure 3, shows how the vowel systems of individuals from each generation change over time. White lozenges indicate the initial vowel system, while the final vowel system is indicated by white squares. It can be observed that the vowel systems contract slightly in acoustic space. However, whereas the five-vowel system remains intact, the seven-vowel system looses the two central vowels.

The way the number of vowels changes over time for seven-vowel systems is presented in figure 6. Again, the population consisted of twenty agents, so in order to calculate the average number of vowels per agent, the numbers on the y-axis have to be divided by 20. It must be stressed that usually, all agents in a population have the same number of vowels. The fact that the graph does not seem to level off on an integer multiple of twenty has to do with the fact that the graph shows the average of a number of populations with different vowel systems.

Sometimes the collapse from a seven-vowel system to a five-vowel system took longer than the 100 generations displayed in figure 5, but in general it can be seen that the number of vowels in the population drops quite quickly, so that within 25 generations or so, most populations will have a five-vowel system. Eventually all populations converge towards a five-vowel system.

The case with agents that learn on the basis of non-reduced speech is comparable but differs in an important respect. As it cannot be expected that speech is produced completely perfectly, a very small reduction of articulation was allowed. The reduction value was set to 0.02. The way vowel systems change over time is illustrated in figure 8. This figure is completely analogous to figures 3 and 5 and follows the development of five- and seven-vowel systems over 100 generations.

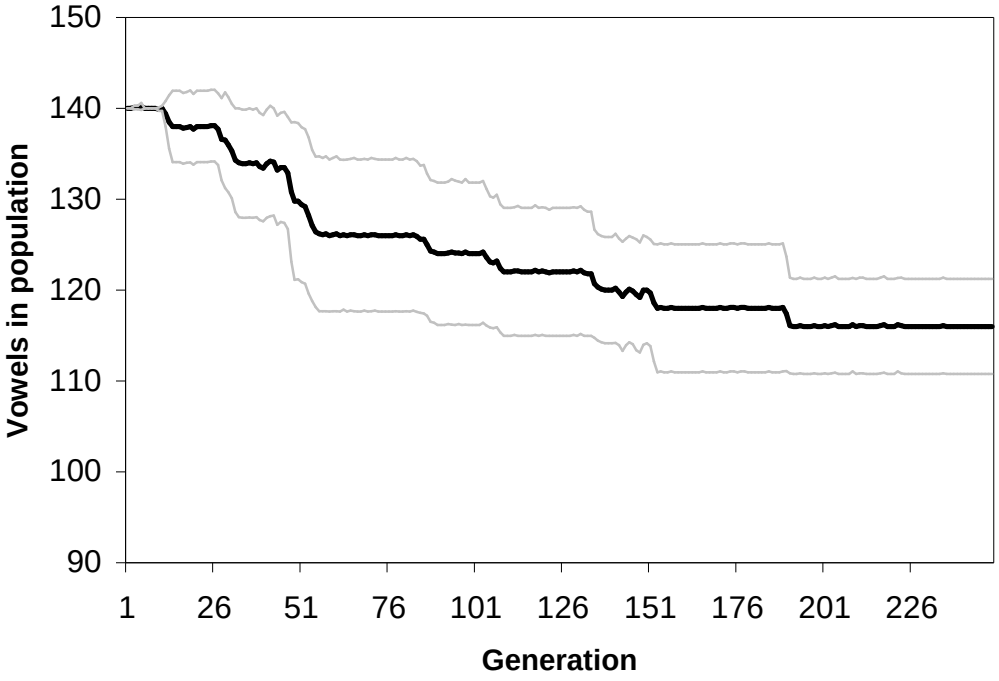

Figure 7: Change over time of the number of vowels in a population of agents where adults use non-reduced speech when addressing infants. Shown are the average over ten runs (bold line) and the 95% confidence interval (grey lines).

Again, the initial vowel system is indicated with white lozenges, while the final vowel system is indicated with white squares.

Although the five-vowel system seems to be slightly better preserved (but this could just be coincidence of the example) the important difference is in the preservation of the seven-vowel system. Whereas in the previous model, all central vowels disappeared quickly from the agent's vowel repertoires, in this model the high central vowel remains in the agents' vowel systems. This turns out not to be a coincidence. Central vowels are preserved better in this model than in the previous model. In figure 7 the way the number of vowels in the population changes over time is shown. It is clear that the number of vowels drops much more gradually than in the previous model. The first central vowel is only lost completely after on average about 50 generations, while the second central vowel seems to be preserved for all 250 generations.

Apparently there is no detectable difference in behavior between the two models for small vowel systems, but for larger vowel systems with central vowels, more careful articulation helps in preserving the vowel system in the population.

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 articulatorily. 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 experiments, two methods of preserving structure in vowel systems were compared. 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 articulatorily by about 20%. However, the infant agents compensated for this by pushing the learned vowel categories away from the center of articulatory space. In the second method, infant agents learned the position of vowel categories on the basis of speech that was only very 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, simple 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 (Labov 1994) 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 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. No conclusion can therefore be drawn about which method best explains learning by infants on the basis of this data. However, a clear difference between the two methods became visible when seven-vowel systems with two central vowels were used. The method in which infant agents compensate automatically for reduced articulation was not able to preserve the central vowels in the vowel system. After approximately 13 generations, half of the populations had already lost both

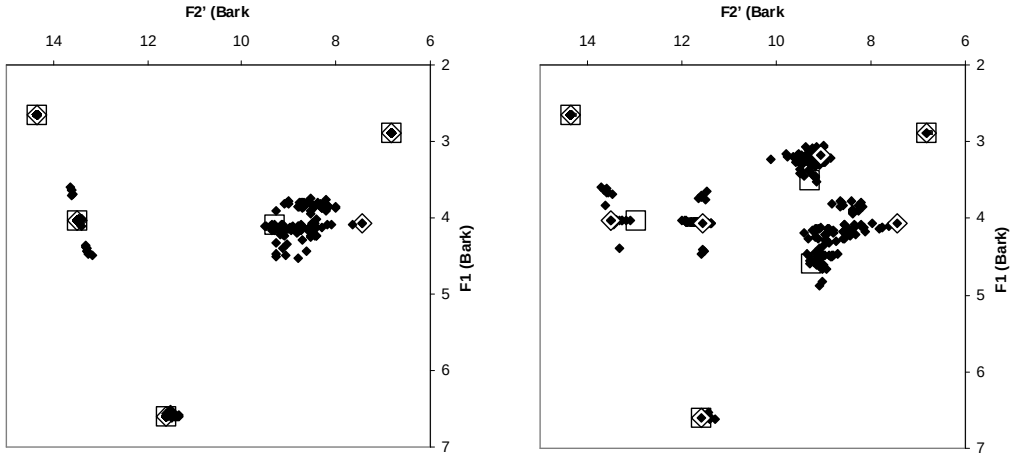

Figure 8: Change over time of vowel systems for the case where infant agents are exposed to non-reduced vowel systems. Shown are 100 generations for an initially five (left) and seven (right) vowel system.

central vowels and all of them had lost at least one central vowel. At the end of the run (250 generations), only two out of ten populations that were investigated still had a central vowel left. On the other hand, the method in which agents learned on the basis of carefully articulated examples was much better able to preserve the central vowels. Only after 53 generations did half of the investigated populations loose at least one central vowel. At the end of the run (also 250 generations), eight out of ten populations still had a central vowel left.

This seems to indicate that large vowel systems with central vowels are more stable when learning takes place on the basis of carefully articulated examples. Learning on the basis of carefully articulated data results in the required stability to account for the way human vowel systems change over time, whereas learning on the basis of reduced data while performing a compensatory expansion does not do so. Although the compensation mechanism used in these experiments was rather crude, and a more sophisticated mechanism could probably 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 (Norwegian in the UPSID database, source: Vanvik 1972) different vowels, and given the experimental results, it is unlikely that such complex systems can be stably reconstructed from reduced articulations.

Apparently, small vowel systems can be transferred from one generation to the next by means of reduced articulation and a simple compensation mechanism. However, larger vowel systems cannot be stably transmitted in this way. For successful transmission of such system, learning needs to be based on more carefully articulated examples. 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 (Kuhl *et al*. 1997) seems to indicate that the more vowels a language has, the more carefully articulated special infant-directed speech is. In this study, words containing the vowels [i], [a] and [u] were recorded from mothers that spoke to their 2 to 5 month old infants (using infant-directed register) and with the experimenter (using adult-directed register). From figure 2 in their paper, 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 (Huei-Mei Liu, *personal communication*) 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 can be used. 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. Apparently, in such cases, infant-directed speech fulfills and adaptive purpose.

Acknowledgements

References

de Boer, Bart (2000) Self-organization in vowel systems, Journal of Phonetics 28(4), pp. 441–465

Fernald, A. (1985) Four month-old infants prefer to listen to motherese. *Infant Behavior and Development* **8**, 181–195

Fukunaga, Keinosuke (1990) *Introduction to statistical pattern recognition*, Boston: Academic Press.

Hock, H. H. (1991) *Principles of Historical Linguistics, second edition*, Berlin: Mouton de Gruyter.

Kirby, Simon (1999) *Function, Selection and Innateness: The emergence of language universals*, Oxford: Oxford University Press.

Kirby, Simon (2002) Natural Language from Artificial Life, *Artificial Life* **8** (2), pp. 185–215

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. (1997) Cross-Language Analysis of Phonetic Units in Language Addressed to Infants, *Science* **277** pp. 684–686

Labov, W. (1994) *Principles of Linguistic Change: internal factors*, Oxford:Blackwell

Ladefoged, Peter & Ian Maddieson (1996) *The Sounds of the World's Languages*, Oxford: Blackwell.

Mantakas, M, J.L. Schwartz & P. Escudier (1986) *Modèle de prédiction du 'deuxiéme formant effectif' F_2'—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, pp. 157–161.

Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée & Christian Abry (1997), The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics* **25**, pp. 255-286.

Steels, Luc (1995) A Self-Organizing Spatial Vocabulary. *Artificial Life* **2**(3), pp. 319–332.

Steels, Luc (1997b) The Synthetic Modelling of Language Origins, *Evolution of Communication* **1**(1): pp. 1–34.

Vanvik, A. 1972. A phonetic-phonemic analysis of Standard Eastern Norwegian. *Norwegian Journal of Linguistics* **26**: pp. 119–64.