HUMAN AND COMPUTER ACQUISITION OF VOWEL CATEGORIES

Bart de Boer
Patricia K. Kuhl
Department of Speech and Hearing Sciences and Center for Mind, Brain and Learning,University of Washington, Seattle, USA

ABSTRACT

This paper describes work in progress on computer modeling of speech acquisition. In these tests, a computer model is presented with natural speech to examine whether the model can learn vowel categories. Preliminary results are promising. The computer learns a three-vowel and a five-vowel system from consonant vowel syllables in isolation. However, the experiments show that clear input is crucial for learning categories and that the most frequent variety of speech, rapid casual speech, does not provide such a signal. We propose that motherese, the speech register that parents use for addressing their children, can provide such an input. The usefulness of motherese in learning vowel categories can be tested with the computer model using existing recordings of motherese and adult-to-adult speech. Computer modeling of this type is compared with speech recognition systems.

1. INTRODUCTION

When one considers speech recognition as a problem of pattern classification, the acquisition of speech becomes a problem of finding the different classes into which the speech signal has to be divided. Work in Kuhl's laboratory [6], [7], [8] and in others (see e.g. Jusczyk [5]) on infant learning has resulted in detailed knowledge of the different stages infants go through when acquiring native-language phonetic categories [8], [9]. Newborn infants are capable of hearing differences between all the phonetic units heard in the languages of the world. As they receive input from their native language however, they lose the ability to distinguish sounds that do not play a role in this language [7]. Infants' perception of native-language speech categories shows a dramatic improvement during the second half of the first year, as shown in recent studies in Kuhl's laboratory [11]. We have argued that the brain's neural commitment to native-language structure causes this change in foreign-language speech perception [9], [10], [11].

In the research described in this paper, we use computer models to investigate how children can acquire the different categories of speech sounds in their native language. The speech sounds investigated here are vowels, as they are easy to detect and represent. We examine the hypothesis that they can acquire vowel systems based purely on the input they receive, without innate specifications of vowel categories. The computer model should show similar learning behavior as the infants if it is a proper model of the acquisition of speech. In exposing computer models to the same input that infants receive, we will test the hypothesis that 'motherese,' with its clearer, slower and simpler speech, aids speech acquisition.

The goal of the work is to examine three central points: (a) that computer models can aid our understanding of humans' acquisition of speech, (b) that speech categories can be acquired through exposure to natural speech, and (c) that motherese plays an important role in speech acquisition.

The computer models can use the know-how that has been gathered by researchers in speech recognition. However, the aim of a model of speech acquisition is different from that of a model of speech recognition. The aim of speech recognition is to build models that perform as well as possible on a given task, without regard of cognitive plausibility. Moreover, a lot of knowledge of the language to be recognized is often built into the speech recognition model, whereas the aim of the speech acquisition model is to learn about the language from scratch. Therefore, existing speech recognition algorithms cannot be used directly, and performance of the model cannot be expected to be as high as that of state-of-the-art speech recognition software. The aim of the model is more to show that speech categories can be learned, rather than to learn them as well as possible.

At the moment, the aim of the computer modeling study is to train the model with speech recordings of women talking to their infants and the same women talking to other adults. It will then be possible to investigate whether motherese is instrumental for learning speech. These experiments are still in progress. The main aim of this paper is to present the learning algorithm and some promising results from training it with real language data.

The next section will present the algorithm in some mathematical detail. This section relies on some concepts from neural network theory and speech signal processing. However, an attempt is made to explain the ideas behind the algorithm in a way that is understandable for researchers with no background in these subjects. Section three presents the details of the experiment and the dataset that was used and points out the ways it can and cannot be extended to recordings of motherese. Section four presents the results and section five discusses them and points out possible avenues of future research.

2. THE LEARNING MECHANISM

The problem that has to be solved by an infant learning the classes of speech sounds in its native language consists of two parts: finding out how many classes there are and finding out where the centers of the classes are located. For the second problem there are standard algorithms. When the number of classes is known, a standard clustering algorithm, such as K-means or hierarchical clustering (for an introduction to clustering in speech processing, see e.g. [2]) can be used. However, an infant cannot know the number of speech sounds in its native language *a priori*. Any algorithm that models the way children learn will therefore have to make an estimate of the number of clusters as well as their positions.

Another prerequisite of the learning algorithm is that it should learn in an incremental way. Infants most probably do not store all the different speech utterances they hear in order to derive the possible speech categories. Rather, their brains are slightly modified each time they hear a sound, until they have learned the sound system of their language. The computer model should behave in a similar way.

There are many possible algorithms that could perform such a task. The one that is proposed here is based on neural network research and was chosen because it is reasonably simple and cognitively plausible. It should be kept in mind that we do not claim that children learn exactly according to this algorithm. The aim of the research described in this paper is just to show that classes of speech sounds can be learned quickly and without innate knowledge using real (though carefully articulated) speech signals.

2.1. Training the Network

The algorithm is a neural network that is based on a Kohonen network [6]. The network consists of a large number of nodes (100 in the experiments presented here, but probably many more are needed for more complex utterances), which are inspired by biological neurons. They have a number of inputs, each of which has a certain weight, and, depending on the input to the network, the nodes become activated, just as happens with real neurons. The nodes compete in a winner-take-all competition so that only the node that has the highest activation remains active. The input weights of this node are then modified slightly so that they move closer to the input signal. The activation of nodes depends not only on the inputs, but also on the frequency of activation: those nodes that have won the competition too frequently in the past have decreased activation, so that other nodes get a chance to win.

The aim of the network is to learn an equiprobable distribution of the nodes. This means that each node has an equal probability of being activated by an input vector, which in turn means that the distribution of the weight vectors follows the distribution of the input vectors. If there are clusters in the inputs, there will be clusters in the network.

Mathematically, the network works as follows. Inputs are vectors $\mathbf{x}$, consisting of elements $x_1 \ldots x_k$ where k is the number of inputs per input vector, equal to two in the experiments described here, but it could be higher for more complex sounds. Each node j has a weight vector $\mathbf{w}_j$ consisting of elements $w_{j1} \ldots w_{jk}$. The activation y_j of each node j can then be calculated as follows:

$$y_j = e^{-0.1\sum_{i=1}^{k}\left(x_i - w_{ji}\right)^2} \qquad (1)$$

The sum in this equation is just the Euclidean distance between the input vector and the output vector. The exponential function causes this distance to behave like a real activation, 1 for minimal distance, indicating maximal activation, approximately linearly proportional to distance for small distances and asymptotically approaching 0 for large distances. The subscript of the most active node will be indicated by m.

For each node, the network also keeps track of a measure (based on [4]) that depends on how often it was activated. This measure is called f_j for node j. It is updated as follows:

$$f_j \leftarrow \begin{cases} (1-\beta)\, f_j & \text{for } j \neq m \\ (1-\beta)\, f_j + \beta & \text{for } j = m \end{cases} \qquad (2)$$

Initially, f_j is given the value of $1/N$ for each node in the network, where N is the number of nodes in the network. The constant β is a parameter of the system and has a value of 0.0001 in all simulations presented here.

On the basis of f, a bias value b is calculated as follows:

$$b_j = \gamma\left(\frac{1}{N} - f_j\right) \qquad (3)$$

where γ is a parameter of the system which has the value 10 for all simulations. The final, most active node is then selected on the basis of the sum of b_j and y_j. The node with the highest value for this sum is said to have the subscript l. The bias term gives an advantage to nodes that are close, but that have not been activated often.

The weights of this node are then updated as follows, in vector notation:

$$\mathbf{w}_l \leftarrow (1-\alpha)\mathbf{w}_l + \alpha\mathbf{x} \qquad (4)$$

where α is a parameter of the system, which has the value of 0.1 for all simulations.

2.2. Testing the network

The aim of the network is not just to learn an equiprobable distribution. In itself this is not very interesting, as an equiprobable distribution can be learned for any input. The point is that the network should show categorization behavior and that the categories should correspond to phonemes in the actual speech signal. Categorization behavior in humans is evident from the fact that different signals are perceived as the same, and that different signals, which are acoustically equidistant, are perceived as the same when they fall within the same category, but are perceived as different when a category boundary passes in between them.

Such categorization behavior can be implemented in a neural network by means of the population vector [3]. A population vector is a weighted average of the weight vectors of the most active nodes of a population after an input has been presented. It turns out that a population vector generally shifts towards the largest density of network nodes.

The population vector is calculated on the basis of the set M of most active nodes of the network. In the simulations presented here, the 12 most active nodes of the network are used. The following formula is used for calculating the population vector **p**:

$$\mathbf{p} = \sum_{j \in M} y_j \mathbf{w}_j \tag{5}$$

The population vector is thus an average that is weighted on the basis of how active the node is.

The point is that the population vector can be iterated (an idea originally due to Pierre-yves Oudeyer of the Sony Computer Science Laboratory in Paris, France [12]) and that it then converges on certain points in the network, which usually happen to be centers of clusters. In this way the network can be used to categorize inputs: an input belongs to the category in which its population vector converges. The convergence point can then be considered the prototype of the category.

2.3. Why this Works

The learning and categorization algorithm works on the basis of two mechanisms: a mechanism to find a representation that reflects the distribution of the inputs (but not as detailed, so that noise is reduced), and a mechanism to find the centers of clusters in this distribution. The actual implementations used here—a neural network and a population vector—are not very important, although they were chosen to be cognitively plausible. Categories of speech sounds are reflected as regularities in the signal. Regularities will form clusters in a well-chosen representation of the input and any mechanism that finds the centers of these clusters will tend to find the categories in the input signal.

It must be noted that the categories are mostly defined by behaviorist principles. Even though it is possible in theory to identify the clusters of nodes in the network, this does not necessarily mean one can find all convergence points. The only way to do this is to present the network with different input signals, and to find the point to which they converge. But this does not contradict what we know about humans. The only way we can learn about the categories of speech sounds in humans is by doing similar experiments. It is unlikely that we would be able to find specific locations where the different categories of speech sounds are stored in the human brain. Representations in neural networks tend to be inherently distributed.

2.4. Processing the Speech Signal

Another important aspect of a computer model is the method by which the signals are processed into a form that the learning mechanism can work with. If the signal processing component produces clean samples, it will be much easier for the learning mechanism to learn something useful—and for the researcher to make sense of what is learned—than if the samples produced by the signal processing component are complicated and noisy.

The signal-processing component used in these experiments extracted the first two formants from the signal. This was done using Linear Predictive Coding-analyses (LPC, [2]) of 256 samples of the input signal (which was sampled ad 11025 Hz). A window was taken for analysis every 64 samples. The LPC used 10 poles. The positions of these poles determine the points of strongest resonance and thereby the formants of the signal. These were found numerically on the basis of the LPC-coefficients.

This analysis works well for clear signals and results in data points that can be represented two-dimensionally, which is an advantage for analyzing and presenting results. However, it is too simple for more complicated and noisy signals. In the discussion we will suggest other possibilities for signal processing.

3. THE DATA SET

The research effort described in this paper was meant to investigate whether the theory underpinning the learning mechanism actually works in practice. It was therefore decided to train the network with a simpler signal than that of the recordings of motherese and adult-to-adult speech. The signals of choice were isolated consonant-vowel syllables uttered by a single male speaker (the first author). Two sets of syllables were generated: one that contained three vowels ([i], [a], [u]) and one that contained five vowels ([i], [ɛ], [a], [ɔ], [u]). Both sets contained the same three consonants ([b], [d], [g]).

The training sets consisted of a random sequence of 400 of these syllables. The signal was sampled at 11025 Hz with 16 bits accuracy, and the volume was scaled such that the average of the signal was 0 and its standard deviation was 2100. From this signal, formants were calculated with the method described above. However, since the aim of the study was to learn vowel categories, formants were only calculated for the most powerful parts of the signal. These parts correspond to the nuclei of the vowels in the syllables and were determined by simple thresholding: only the parts of the signal were analyzed where the sum of the power (calculated as the square of the amplitude) over 256 samples was greater than 10^8. Other parts of the signals were discarded. The formants were first calculated in Hertz and then converted into Barks (a perceptually motivated frequency scale) with the following formula:

$$B = 6 \cdot \ln\left(\frac{H}{600} + \sqrt{\left(\frac{H}{600}\right)^2 + 1}\right) \tag{6}$$

where H is the frequency in Hertz and B is the frequency in Barks.

Automatic formant analysis as used in this research sometimes tends to produce formant peaks in the wrong places. For this reason, signals with unrealistic formant frequencies (lower than 2 Barks for the first formant and

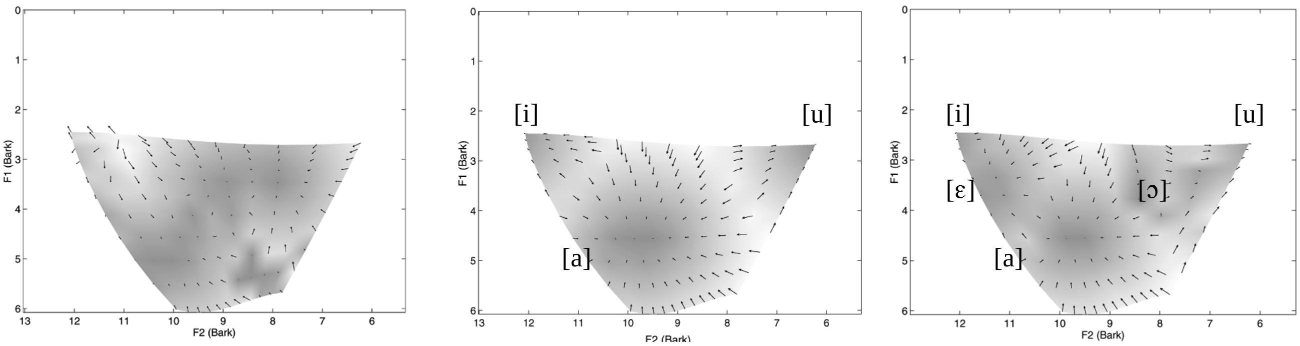

Figure 1: The results of learning to classify three- and five-vowel systems. The leftmost frame shows classification behavior without training. The middle frame shows classification behavior after exposure to a three-vowel system and the rightmost frame shows classification behavior after exposure to a five-vowel system.

lower than 4.5 Bark for the second formant) were discarded, too.

These formant values were stored in a file and while the network was trained, the training samples were taken from this file in sequence. When the end of the file was reached, training samples were again taken from the start of the file, such that the whole file was presented several times.

Test formant pairs were generated with an artificial formant synthesizer, described in [1]. The articulatory parameters of this synthesizer were changed from 0 to 1 in increments of 0.1 for both tongue position (in the front-back dimension) and tongue height. Rounding was defined to be equal to position so that front vowels were unrounded and back vowels were rounded. In total, 121 test formant pairs were generated.

4. RESULTS

The results are summarized in figure 1. In this figure, the classification behavior for the test data set is shown (from left to right) for an untrained network, a network that was trained with a three-vowel system and a network that was trained with a five-vowel system. The number of training examples presented to the network was 300,000, corresponding to roughly 20 hours of speech. Classification behavior in these figures is indicated in two different ways. The first uses arrows that start at the original test data point and point in the direction of the point to which they converged after 20 iterations. The length of the arrow is proportional to the distance of the displacement of the iterated population vector. The arrows therefore point towards the points of convergence. The second method uses color to indicates where the points of convergence are. The darker the color, the less the population vector was displaced. The darkest spots are therefore the points of convergence.

The axes give frequencies in Barks. The orientation of the axes is inverted with respect to their default orientation in order to project the vowels such that their positions will be most familiar to phoneticians: high front vowels appear in the upper left corner and low vowels in the lower part of the graph. Due to articulatory restrictions, not all arbitrary pairs of first and second formant value can be used, but only a roughly trapezoidal area is accessible (shown in gray in the graphs).

It is clear from these graphs that categorization behavior occurs that corresponds to the vowel system to which the network was exposed. The untrained network shows classification-like behavior, but there are many points of convergence, some of which apparently fall outside the area for vowels (judging from the arrows that point away from the accessible acoustic space). Also, the points of convergence are not located at places that would be expected in a human vowel system.

After exposure to a three-vowel system, the network clearly shows three points of convergence. These correspond to [i], [a] and [u], as indicated in the graph. The basins of attraction (the sets of starting that are attracted to a given convergence point) are not quite realistic. A human subject would be more likely to classify any high vowel (on a straight line between [i] and [u]) as either of these two, but the system classifies certain high vowels as [a]. This can probably be solved by weighting the second formant properly with respect to the first formant. Studies of modeling vowel systems [13] have shown that the distance in the second formant dimension should be considered approximately 30% as important as distance in the first formant dimension. This would put [i] and [u] much closer together, and would probably attract any signal between these two categories towards one of them.

The case of the five-vowel system is also clear, but the positions and relative sizes of the convergence points are somewhat unexpected. The dark areas and arrows indicate clear areas of convergence for vowels [i], [ɛ] and [a]. The situation for the back rounded vowels [ɔ] and [u] is slightly less clear. There seems to be quite a large area where the network seems to have trouble deciding whether a signal belongs to category [ɔ] or category [u]. If one follows the arrows, however, one finds that there is a convergence point for both categories. Again, the basins of attraction for the different data points are not quite realistic, but the system does show clear classification behavior, and the number and approximate position of the convergence points corresponds to the number and position of the vowels in the data set.

5. DISCUSSION

These results show that it is possible to learn categories of speech sounds from a sufficiently clear, but real signal in an unsupervised way. It was shown that a simple neural network could learn the vowel categories from a signal consisting of consonant-vowel syllables, in which all the consonants are voiced plosives. This indicates that in principle children could do the same. The task facing an infant is much more complex, since the sound systems of human languages are much more complex. Realistic sound systems also contain other speech sounds, such as voiceless obstruents, nasals, liquids and, quite often, more complex consonants, that require a combination of articulator movements for their production. In a completely realistic system, these speech sounds would also have to be learned, but since the focus of this research is on vowel systems, the problem here is how to extract the vowels from these more complex sounds. This probably requires more complex signal processing and a more complex representation of signals.

Moreover, the signals that were learned here were produced by a single speaker. Infants are probably exposed to input of multiple speakers, although there are some indications that they tend to focus on a single speaker (particularly their mother) for certain cues, such as intonation. The categories learned by this model are still quite speaker dependent, as can be seen from the position of the [a] category in figure 1. The second formant of the [a] in the training set was much higher than that of the [a] in the test set, and the convergence point for [a] is therefore higher than expected.

A more sophisticated representation (one based on spectra that are weighted in a psychologically realistic way is under development) would probably address some of these problems but would make it harder to represent and interpret the results. Right now, both the training data set and the test data set contain two-dimensional points. These are easy to represent in figures. The weighted spectra would require at least 16-dimensional data sets, which would make it much harder to investigate and represent what the network has learned.

Preliminary experiments with signals from mothers talking to their children and talking to other adults (also used in the work described in [8]) have been done. These show convergence, but so far it has been impossible to ascertain whether the convergence points correspond to the vowels in the input data. Thus it has not been possible to check whether motherese input facilitates learning or not.

However, the ease with which experiments can be done with the computer model and the ability it has shown to learn vowel categories promise interesting results for this method of investigation.

6. ACKNOWLEDGEMENTS

This research was performed at the Center for Mind, Brain, and Learning of the University of Washington. It was supported by grants from the National Science Foundation and the Human Frontiers Science Program and by research support for the Center for Mind, Brain and Learning from the Talaris Research Institute of Seattle, Washington. Correspondence should be directed to: Bart de Boer, University of Washington, Center for Mind, Brain, and Learning, Box 357988, Seattle WA 98195, USA. Email: bdb@u.washington.edu.

7. REFERENCES

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

[2] Deller, J.R., Hansen, J. H. L. & Proakis, J. G. (2000). *Discrete-Time Processing of Speech Signals*. New York: IEEE Press.

[3] Guenther, Frank H. and Gjaja, Marin N. (1996). The Perceptual Magnet Effect as an Emergent Property of Neural Map Formation. *Journal of the Acoustical Society of America*, **100**, 1111–1121

[4] Hecht-Nielsen, Robert, (1990). *Neurocomputing*. Reading, MA: Addison-Wesley.

[5] Jusczyk, P. W. (1997) *The Discovery of Spoken Language*, Cambridge (MA): MIT Press.

[6] Kohonen, Teuvo, (1988). The “Neural” Phonetic Typewriter. *IEEE Computer*, **21**(3), 11–22

[7] Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N. and Lidblom, B. Linguistic experience alters phonetic perception in infants by 6 months of age, Science **255**, pp. 606–608

[8] Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants, *Science*, **277**, 684-686.

[9] Kuhl, P.K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Science*. **97**(22), 11850–11857.

[10] Kuhl, P.K. (2000). Language, mind, and brain: Experience alters perception. In M.S. Gazzaniga (Ed.), *The new cognitive neurosciences (2nd edition).* Cambridge, MA: MIT Press. pp. 99-115

[11] Kuhl, P.K., Tsao, F. M., Liu, H. M., Zhang, Y. & de Boer, B. (*in Press*) Language/Culture/Mind/Brain: Progress at the Margins between Disciplines, *Annals of the New York Academy of Sciences*.

[12] Ouderyer, P-y, Coupled Neural Maps for the Origins of Vowel Systems, to be presented at *International Conference on artificial Neural Networks*, 2001

[13] Schwartz, J. L., Boë, L. J., Vallée N. & Abry, C. (1997). The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics* **25**, 255-28.