



three target vowels /i/, /a/, and /u/ in the data sets. After learning is complete, the learned formant values are compared with standard formant values of the target vowels, and thus give an estimate of the quality of the learned vowel system.

We do not claim that the computer model is a direct model of the way infants learn vowels, or that the signal processing algorithms that are used are as sophisticated as the signal processing abilities of infants. However, the aim of the study is to compare the learnability of the ID and AD data sets. Because the computer model is the same for both data sets, the comparison is fair.

## 2. Data set

The data set was based on recordings of mothers talking to their infants and of the same mothers talking to an adult [Gustafson, 1993; Kuhl *et al.*, 1997]. The mothers were engaged in conversation by the experimenter about a number of objects and were asked to play with and talk to their infants using the same objects. The objects (i.e., bead, pot, boot, sheep, sock, shoe) were selected to contain the American English vowels /i/, /a/, and /u/. Words for the objects were isolated from the recordings.

For this study, three words were used: sheep, sock, and shoe. Sheep, sock, and shoe were selected from all target words, because the voiceless fricatives were easiest to distinguish from the vowels. Although shoe and sheep contained a fair amount of co-articulation, (due to the palatal fricative) this turned out not to confuse the computer model, even though an implicit assumption was that vowels are static. However, in these words, /i/ and /u/ tended to be centralized, and thus slightly closer together than in other contexts. Data from ten mothers were used. The number of tokens for each word and each register for each mother is given in table 1.

The words were digitally sampled from cassette tape at 16 kHz with 16-bit accuracy. Signal-to-noise ratio was estimated to be about 40–45 dB.

**Table 1. Number of tokens (and original formant pairs, see § 3.1) per word, register, and mother**

mother	Adult-directed			Infant-directed		
	sheep	sock	shoe	sheep	sock	Shoe
AG	<b>4</b> (9412)	<b>2</b> (9304)	<b>5</b> (20 539)	<b>6</b> (30 716)	<b>4</b> (22 593)	<b>3</b> (18 866)
AH	<b>6</b> (14 029)	<b>5</b> (15 643)	<b>9</b> (37 117)	<b>6</b> (24 967)	<b>9</b> (35 723)	<b>7</b> (22 543)
AL	<b>8</b> (18 806)	<b>3</b> (6921)	<b>9</b> (32 997)	<b>9</b> (38 126)	<b>7</b> (40 196)	<b>8</b> (27 384)
AO	<b>4</b> (7941)	<b>3</b> (12 414)	<b>3</b> (6441)	<b>9</b> (27 756)	<b>6</b> (19 736)	<b>3</b> (25 905)
AP	<b>8</b> (29 513)	<b>6</b> (22 767)	<b>4</b> (10 110)	<b>7</b> (30 869)	<b>9</b> (41 406)	<b>6</b> (40 018)
AS	<b>7</b> (19 916)	<b>8</b> (28 359)	<b>7</b> (21 633)	<b>7</b> (31 137)	<b>7</b> (21 619)	<b>6</b> (35 546)
AT	<b>3</b> (9420)	<b>3</b> (10 477)	<b>3</b> (8499)	<b>5</b> (12 121)	<b>7</b> (54 130)	<b>4</b> (27 386)
AW	<b>8</b> (16 443)	<b>4</b> (12 109)	<b>4</b> (10 754)	<b>8</b> (33 268)	<b>6</b> (35 561)	<b>5</b> (27 124)
AX	<b>4</b> (15 838)	<b>7</b> (34 152)	<b>7</b> (20 083)	<b>8</b> (41 969)	<b>7</b> (29 949)	<b>5</b> (17 057)
AZ	<b>4</b> (11 965)	<b>6</b> (22 971)	<b>9</b> (20 450)	<b>4</b> (16 890)	<b>7</b> (35 663)	<b>6</b> (34 239)

## 3. Computer model

The aim of the study is to compare the learnability of two data sets, rather than to make an accurate model of the human brain. A standard technique from statistical pattern recognition for learning to classify data sets, expectation maximization (EM) of a mixture of Gaussians [Bilmes, 1998; Dempster *et al.*, 1977] was therefore adopted. This technique is fast and easy to re-implement. Its operation is more transparent than that of biologically inspired neural networks but it has parallels with these techniques [e.g., Specht, 1990]. One could raise the objection that a more cognitively accurate technique should have been used, if one wants to say something about learnability for humans. However it is unclear what technique to use. A neural network for unsupervised classification could do the task, but these models are not generally based on an accurate model of the human brain, either. They would introduce an unknown learning bias, and their operation would not be transparent. Also, the performance of various computational learning techniques on the relatively

simple task of classifying a data set into a known number of clusters does not differ much, so it was decided to use the simplest and most transparent learning algorithm. EM in our experiment faces an important part of the learning task that infants face, but it is less powerful than human learning, and it is therefore safe to assume that a data set that results in better learning by EM, results in better learning by humans.

The model has to learn the qualities of the vowels that occur in the speech signal that it receives as input. This means that it has to determine the shape of the distribution of each of the vowel categories in acoustic space. In solving this task it is assumed that the number of categories is known. We do not hypothesize that infants "know" the number of vowels in their language beforehand. We make the assumption because unsupervised classification of a data set is much easier when the number of categories is known beforehand. Very few stable algorithms exist for unsupervised classification with an unknown number of clusters. The assumption that the number of categories is known beforehand is therefore a temporary solution adopted for comparing the two data sets. Learning the number of categories remains a task for later research. Studies of infant perception demonstrate that infants initially discriminate between the phonetic units of all languages, and this essentially partitions the acoustic space into crude phonetic categories [see Kuhl, 2000; Kuhl *et al.*, 2001 for discussion]. It is possible that infants use this ability to make an initial estimate of the number of categories of speech sounds.

### 3.1. Signal-processing

The signal-processing procedures were meant to automatically extract formants from the voiced parts of the target words. The procedures are standard signal-processing procedures, whose parameters were optimized for the task at hand. This means that some of the parameter settings might not be optimal for general formant extraction.

The words contained only one syllable, so vowels were detected because they were voiced. Voicing was determined by the zero-crossing rate. To calculate this, the signal was center-clipped. All samples with an absolute value of less than 50 were set to zero; other samples kept their original value. If the center-clipped signal had fewer than 40 zero-crossings (in both directions) per window of 256 samples (16 ms), the signal was considered voiced, unless the maximum absolute value of all samples was less than 100, in which case it was considered as silence.

After determining the voiced parts of the signal, formants were calculated for windows of 256 samples. These windows were shifted by 1-sample increments over the duration of the vowel, so, effectively for every sample in the vowel, a formant pair was calculated. For each vowel, this procedure results in a formant track, rather than a single formant pair, thus representing it more accurately.

The signal was pre-emphasized by taking the difference between subsequent samples:  $s'_i = s_i - s_{i-1}$ . This emphasizes high frequencies, which is useful for formant extraction. To reduce edge effects, each sample  $s_i$  ( $1 \leq i < 256$ ) in the window was multiplied by  $(i-1) \cdot (256-i)$  (a parabola that is zero at the edges with a maximum in the middle). This makes formant peaks sharper (see, *e.g.*, Press *et al.*, 1992, fig. 13.4.2).

After this preliminary processing, 18 LPC-coefficients were calculated using the recursive Burg algorithm (as described in Press *et al.*, 1992, ch. 13.6). On the basis of these LPC-coefficients, the polynomial that describes the frequency response of the vocal tract was determined and its roots were approximated numerically. The 17 roots ( $z_1, \dots, z_{17}$ ) represent the resonance frequencies (formants) of the vocal tract and their bandwidths. The relation is as follows: *formant frequency* =  $\text{angle}(z_i) F_{\text{sample}} / 2\pi$  and *bandwidth* =  $F_{\text{sample}} (1 - |z_i|) / \pi$ , where  $F_{\text{sample}}$  is the sampling frequency, and  $\text{angle}(z_i)$  is the angle of complex value  $z_i$ . These formulas were taken from (Allen *et al.*, 1987).

Formants were rejected if their frequency was lower than 100 Hz or if their bandwidth was higher than 500 Hz. This rejected spurious peaks in the formant pattern. Also, the first formant frequency had to be lower than 1500 Hz, and the second formant frequency had to be lower than 3000 Hz. This rejected parts of the signal in which the first or second formant could not be found.

The number of formant pairs calculated for different vowels per mother differed considerably (Table 1) because there was variation in the number of tokens per vowel and their lengths. Longer vowels tend to be articulated better than shorter ones. It can indeed be observed in the results that the shortest vowel /a/ is the most dispersed. However, this effect is present in both ID and AD speech, so it does not matter for the comparison. To not bias learning towards vowels with large numbers of formant pairs, it was decided to randomly select an equal number of 1000 formant pairs per vowel, mother and register. These formant pairs were selected with replacement. Thus, for each mother and speech register, the training data consisted of 3000 formant pairs, 1000 each per vowel. Other formant pairs were ignored, but due to high correlation between successive formant pairs, little information about the vowels' average positions was lost.

### 3.2. Fitting a mixture of Gaussians

To determine the learnability of a data set, a mixture of three Gaussians was fitted, and it was checked how close the means (centers) of the Gaussians were to estimated standard formant values (see Fig. 1) of the vowels /i/, /a/, and /u/.

A probability distribution that consists of a mixture of Gaussians is the weighted sum of a fixed number of ordinary (possibly multidimensional) Gaussian distributions. Each individual Gaussian distribution is characterized by its mean and its covariance matrix. The weights of the Gaussians have to sum to one. The weights, the means, and the covariance matrices are the *parameters* of the Gaussian distribution. They have to be estimated on the basis of the data. A well-known algorithm for this is the expectation maximization (EM) algorithm [Bilmes, 1998; Dempster *et al.*, 1977]. On the basis of an initial guess of the parameter values and the observed data, a new, better estimate of the parameter values can be calculated. These steps are iterated until no more improvement is achieved. Because this is difficult to measure in practice, the system was iterated 70 times, after which parameter values stabilised. The system does not necessarily converge towards the best solution. Often it converges to a local optimum, which is less good than the global optimum. The solution found depends strongly on the exact distribution of the data points. Two data sets with similar means and variances can therefore show quite different learnability, as the learning process is influenced by outliers or overlap between vowel categories, among other factors.

The mixture used here consisted of three two-dimensional Gaussians. The parameters were initialized as follows: the weights were set equal, whereas the covariance matrices were set to the unity matrix times 1000. The starting averages were set to the points (400, 2500), (800, 1500) and (400, 800). These roughly correspond to the corners of the acoustic vowel space of a female speaker. This was done to give the system as large a likelihood as possible to generate a mixture in which each Gaussian corresponds to a vowel.

Although learning a mixture of Gaussians can in a sense be considered curve fitting, the approach is more informative than a straightforward statistical analysis of the clusters. The quality of what is learned depends the exact structure of a data set and on the way the learned representation changes over time. The performance of a learning mechanism therefore gives a better measure of the learnability of a data set than a statistical analysis in terms of averages alone (such as was used by Kuhl *et al.*, 1997) or even one using covariance matrices and other higher order moments.

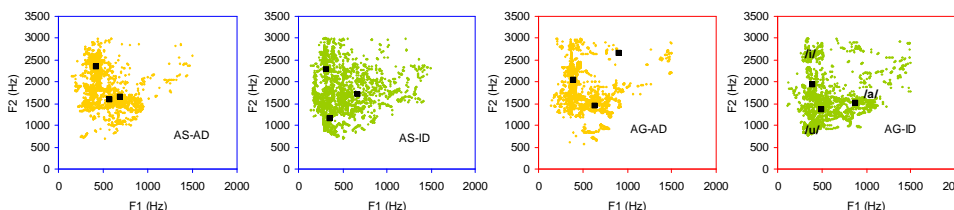


Fig. 1. Comparison of AD speech (yellow dots) and ID speech (green dots) learnability. The black points indicate the centers of the learned categories. Data of mothers AS (blue, left) and AG (red, right) are shown. Note that the ID-based category centers correspond better with actual /i/, /a/, and /u/ (approximate position shown in right graph).

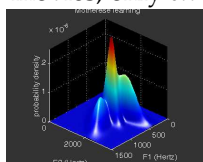
**4. Results**

The means of the Gaussians that have been fitted to infant-directed data correspond better to the original vowels in the input than for adult-directed data. This phenomenon is illustrated in Fig. 1. This figure shows all 3000 data points (3 words times 1000 data points) for both infant-directed and adult-directed registers for mothers AG and AS. Other mothers are similar and can be found in the larger picture.

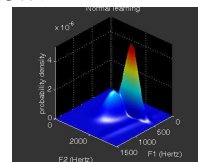
Mm. 3. Picture of all 10 learned systems

It is clear that, like prototypical vowels, the infant-directed tokens are spread out more through the acoustic space (conforming to the findings of [Kuhl *et al.*, 1997]). However, the area per vowel in the acoustic space are larger, so overlap may be greater and it is not directly clear from which of the two data sets the best vowel representations can be learned. But the results show that the means of the Gaussians that make up the mixture (indicated with black squares in the figures) correspond better to the original three vowels (/i/, /a/, /u/) in the infant-directed cases than in the adult-directed cases. The adult-directed cases either have an unrealistic outlier ( $F_1 \approx 1000$  Hz,  $F_2 \approx 2500$  Hz) or show only two distinct peaks (two of the three Gaussians mostly overlap). In the cases in which both have three peaks in approximately the right places, the infant-directed data will invariably have vowels that are closer to the standard vowels. There is no exception to this pattern for the ten mothers tested, indicating that this is a statistically significant finding (assuming equal learnability, this finding would occur by chance with  $p < 0.01$ ).

The learning system very quickly converges to a solution. This is illustrated in the following movies, which show learning based on both AD speech and ID speech for mothers AG and AS. It can be seen that the systems that are learned on the basis of ID speech show the clearest peaks that correspond best to the target vowels. Note that initially only one peak appears, as they all overlap. Also in some of the movies, only two peaks appear to emerge, as the third peak is very low.



- Mm. 4. Movie of learning AG’s ID speech. (839K)
- Mm. 5. Movie of learning AG’s AD speech. (742K)
- Mm. 6. Movie of learning AS’s ID speech. (866K)
- Mm. 7. Movie of learning AS’s AD speech. (730K)



The general conclusion is that ID speech contains better structure for learning a classification. This supports the hypothesis that ID speech may help learning.

**5. Conclusion and discussion**

The results have shown that, given the number of vowels in the vowel system to be learned, their positions can be learned more accurately on the basis of ID speech

recordings than on the basis of AD speech recordings. This indicates that vowel quality is more learnable on the basis of ID speech than on the basis of AD speech. This supports the hypothesis that the modifications that are made to ID speech potentially serve an adaptive purpose. Although the learning algorithm, expectation maximization of a mixture of Gaussians, does not directly model the way the human brain works, it does allow for a comparison of the different data sets, because the learning algorithm and the signal processing routines are the same for both cases, and they are not biased. Future research will focus on the properties that make a data set learnable when the number of categories is unknown. An unsupervised clustering algorithm (such as the one described in Fukunaga, 1990) needs to be used in this case.

This research is the beginning of the investigation of infant acquisition of speech and language using computer models. The result so far has been modest: it has shown that if the number of vowel categories is known, vowels are more learnable from infant-directed speech than from adult-directed speech. An important contribution is also that computer models can be used successfully to investigate developmental and cognitive questions.

### Acknowledgments

The research reported here was supported by grants to Patricia K. Kuhl from NIH (HD37954) and the Human Frontiers Science Program (RG0159), as well as from the Talaris Research Institute and Apex Foundation, the family foundation of Bruce and Jolene McCaw. The authors thank Jean Andruski for providing advice on the measurement and theoretical issues addressed here.

### References and links

- Allen, J. M., Hunnicutt, S., and Klatt, D. (1987). *From text to speech: The MITalk system* (Cambridge University Press, Cambridge).
- Bilmes, J. A. (1998). *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models* (U.C. Berkeley technical report TR-97-021). <http://www.cs.ucr.edu/~stelo/cs260/bilmes98gentle.pdf>
- Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society series B* **39**, 1–38.
- Fernald, A. (1985). "Four month-old infants prefer to listen to motherese," *Infant Behavior and Development* **8**, 181–195.
- Fernald, A., and Kuhl, P. (1987). "Acoustic determinants of infant preference for Motherese speech," *Infant Behavior and Development* **10**, 279–293.
- Ferguson, C. A. (1964). "Baby talk in six languages," *American Anthropologist* **66** (6 part 2) 103–114.
- Fukunaga, Keinosuke, (1990). *Statistical pattern recognition, second edition* (Morgan Kaufmann, San Diego).
- Gustafson, K. T. (1993). *The Effect of Motherese Versus Adult-Directed Speech on Goodness Ratings of the Vowel /i/* (Master of science thesis, University of Washington).
- Kuhl, P. K. (2000). "A new view of language acquisition," *Proceedings of the National Academy of Sciences*, **97**, 11850–11857. <http://www.pnas.org/cgi/reprint/97/22/11850.pdf>
- Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). "Linguistic experience alters phonetic perception in infants by 6 months of age," *Science* **255**, 606–608.
- Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (1997). "Cross-language analysis of phonetic units in language addressed to infants," *Science* **277**, 684–686.
- Kuhl, P. K., Tsao, F. M., Liu, H. M., Zhang, Y., and de Boer, B. (2001). "Language/culture/mind/brain: Progress at the Margins Between Disciplines," in *Unity of knowledge: The convergence of natural and human science* edited by A. Domasio et al. (The New York Academy of Sciences, New York) 136–174.
- Liu, H.-M., Tsao, F.-M., and Kuhl, P. K. (2000). "Support for an expanded vowel triangle in Mandarin motherese," *International Journal of Psychology* **35**(3–4) 337
- Liu, H.-M., Kuhl, P. K., and Tsao, F.-M. (2003). "An association between mothers' speech clarity and infants' speech discrimination skills," *Developmental Science* **6**(3) F1–F10.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C: The art of Scientific Computing second edition*, (Cambridge University Press, Cambridge).
- Snow, C. E. and Ferguson, C. A. (eds.) (1977) *Talking to Children: Language Input and Acquisition*, Cambridge: Cambridge University Press.
- Specht, D. (1990) Probabilistic Neural Networks, *Neural Networks* **3**(1) 109–118