Ecological Language Acquisition via Incremental Model-Based Clustering

Giampiero Salvi

KTH (Royal Institute of Technology),
Department of Speech, Music and Hearing,
Stockholm, Sweden
giampi@kth.se

Abstract

We analyse the behaviour of Incremental Model-Based Clustering on child-directed speech data, and suggest a possible use of this method to describe the acquisition of phonetic classes by an infant. The effects of two factors are analysed, namely the number of coefficients describing the speech signal, and the frame length of the incremental clustering procedure. The results show that, although the number of predicted clusters vary in different conditions, the classifications obtained are essentially consistent. Different classifications were compared using the *variation of information* measure.

1. Introduction

One of the aims of the project MILLE [1] is to analyse the interaction between infants and their linguistic environment in order to model the language acquisition process.

According to the ecological theory of language acquisition [2], the infant is phonetically and linguistically naïve. One of the challenges is therefore the analysis of emergency of phonetic classes in the presence of linguistic stimuli.

As well known by the speech signal processing and automatic speech recognition communities, modelling speech may be seen as a two-fold problem, as not only the acoustic characteristics of speech sounds are of interest, but also their evolution and interaction in time.

Semi-supervised learning techniques [3, 4] have been employed in the past in the attempt to optimise acoustic units and lexica for automatic speech recognition (ASR) tasks, or to find the best acoustic model topology [5]. In the study of time series, clustering techniques have been used in the context of Markov chains in order to classify fixed [6, 7] or variable [8, 9] length sequences.

In this study we focus on the static problem; the unsupervised classification of speech sounds according to their spectral characteristics, using clustering methods. The problem of integrating the sounds into longer sequences, such as syllables or words, is left for future research. Note however, that the two problems are strongly interconnected.

The aim of this study is not to model the psycholinguistic processes taking place during learning in details, but rather to explore the linguistically relevant acoustic environment the infant is exposed to with un-supervised learning techniques.

One of our concerns was modelling the phonetic acquisition process *incrementally* both in the attempt to mimic the intrinsic incremental nature of learning and because of the clustering methods limitations with large datasets. This is in agreement with studies on perception that investigate the properties of acoustic memory and of the stores we can rely on in order to analyse and recognise sounds [10].

2. Method

2.1. Clustering and parameter estimation

Model-based clustering [11, 12, 13] is among the most successful and better understood clustering methods. This is a parametric procedure that assumes that the data points are generated by a mixture model with density

$$\prod_{i=1}^{n}\sum_{k=1}^{G}\tau_k f_k(x_i|\theta_k)$$

Where τ_k and θ_k are the model parameters and $f(x_i|\theta_k)$ is a probability distribution. In our case the shape of each distribution is assumed to be Normal and its parameters are the means μ_k and covariances Σ_k. Furthermore we assume the covariance matrices to be diagonal, with ellipsoidal shape and varying volume across Gaussian components.

A common procedure for finding the model that best fits the data is to use model based hierarchical clustering [12, 14, 15] as an initialisation procedure. The EM algorithm [16] is then used to fit the mixture model parameters with a fixed number of components G. Both the distribution form and the value of G can be varied in order to obtain models of different complexity. The Bayes information criterion (BIC) [13], defined as

$$BIC \equiv 2l_M(x,\theta) - m_M \log(n)$$

is finally used to select between different models, in the attempt to find a trade-off between the model fit to the data (likelihood $l_M(x,\theta)$), the model complexity in terms of number of independent parameters m_M and the amount of available data points n to estimate the model parameters.

With our choice of distribution form, the complexity of the model is controlled exclusively by the parameter G, that corresponds to the number of classes.

Recently Fraley et. al. [17] introduced an incremental procedure to overcome the problem of model initialisation with large datasets. The procedure obtains an initial mixture model on a subset of the data. New data is matched against the current model and the data points are divided into two groups, A and B, depending on how well they fit the current representation. A new model is initialised using the current classification for the points in group A and a new class for the ones in group B. Eventually the procedure is iterated to find the best number of mixture components. The BIC is used at each step to select the most parsimonious model that best fits the data. In Fraley's examples the data points have no specific order, and the data subsets are sampled randomly.

In this study we employ a similar procedure, where the new data is fed into the system in successive time ordered frames.

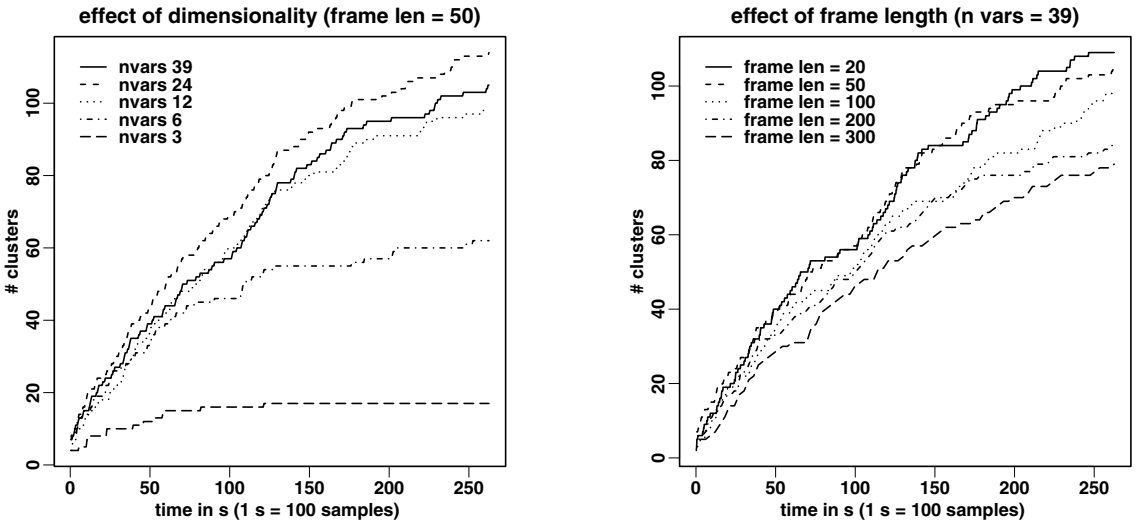

Figure 1: Number of clusters for each iteration and for varying number of coefficients (left) and frame lengths (right)

frame len (sec)	# of coefficients	# clusters	final BIC
50 (0.5)	3	17	-446426.7
”	6	62	-870400.2
”	12	98	-1451219
”	24	114	-2896358
”	39	105	-4127347

frame len (sec)	# of coefficients	# clusters	final BIC
20 (0.2)	39	109	-4123062
50 (0.5)	”	105	-4127347
100 (1)	”	98	-4116574
200 (2)	”	84	-4120498
300 (3)	”	79	-4107472

Table 1: Number of clusters and BIC value for the final models with varying number of coefficients (left) and frame lengths (right)

One difference with Fraley's study is that, given the sequential nature of our problem, we can expect the new data points to follow a regular evolution in time. We are thus interested not only in the final model, but also in the way this model evolves in time.

A limitation with this method is that the number of classes can only increase during the process, while it is possible that a more complete sample of the data would reveal a simpler structure. This problem is somewhat more evident in our case as the subsets of data are not randomly chosen.

The incremental model based clustering procedure was implemented by the author in the statistical program `R` [18] relying on the implementation of the EM and the model-based hierarchical clustering algorithms from the `MCLUST` [19] package.

2.2. Evaluation

Given the task of this investigation, it is not easy to establish a procedure to evaluate the results. The first difficulty is defining what should the target classes be, as it is not known how these classes evolve in time in the infant.

Secondly, the optimal choice for acoustic classes is strongly dependent on the task of discriminating meanings, which involves higher level processing and time integration that are not considered in this study.

Moreover, from an information theoretical point of view, it is not clear that a model that optimally represents the acoustic properties of speech sounds should correspond to a phonetic classification. In ASR for example, each phoneme is modelled by a large number of distributions to represent its variation with contextual variables.

In the absence of a good reference, we concentrate at this stage on evaluating the consistency across classifications in different conditions, in an attempt to highlight possible sources of errors due to limitations of the methods.

A measure of consistency is given in [20] and relies on information theoretical concepts. The so called *variation of information* defined as the sum of the conditional entropies of clustering C given clustering C' (and vice-versa), forms a metric in the space of possible clusterings, and assumes the value 0 for identical classifications. This was taken as a measure of disagreement between C and C'.

Finally some examples on how the clusters evolve in time are given together with a spectrogram in order to compare the emergent classes with acoustical features.

3. Experiments

3.1. Data

The data used in this study is an excerpt from the recordings made at the Department of Linguistics at Stockholm University [2, 21, 22]. The interactions between mothers and children are recorded with microphones and video cameras. Most of the acoustic data consists of child-directed speech by the mother. As a reference, the mothers are also recorded when speaking with an adult.

A twelve minutes recording of the interaction between a mother and her child was selected. Only the child directed acoustic data was used. Pauses and segments with the infant's voice were removed before processing. From the sound, thirteen Mel frequency cepstral coefficients (including the zeroth)

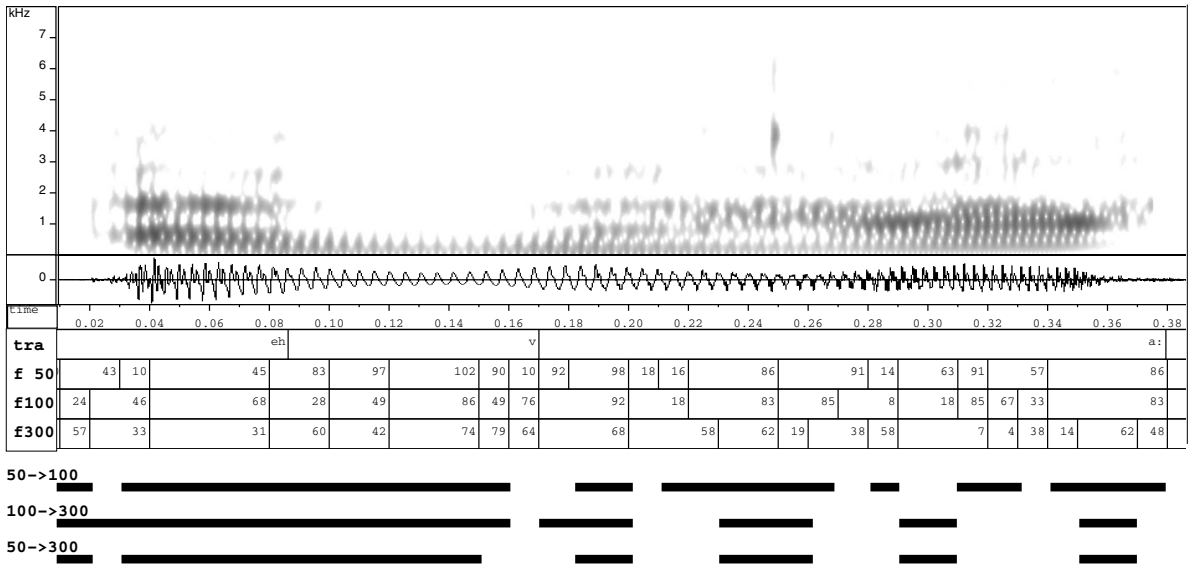

Figure 2: Example of classifications with different frame lengths. The phrase contains "eh, vad" (eh, what). The transcriptions are respectively a reference, the classification with frame length 50, 100 and 300 samples. Note that the class numbers are mere labels and are randomly assigned by the clustering procedure. The thick line represent the agreement between classifications pairwise.

were extracted at 10 msec spaced frames, and their differences were computed up to the second order, for a total of 26254 vectors with 39 coefficients. The total set of parameters is $\{c_0 - c_{12}, d_0 - d_{12}, a_0 - a_{12}\}$ where c are the static coefficients, d the differences and a the second order differences.

3.2. Experimental settings

Two of the factors involved in the process have been investigated in the experiments.

The dimensionality of the data has an interesting interpretative value in this context as the classification obtained can be interpreted in the light of the discriminative features that the infant can rely on. This factor has also a technical importance as all statistical and machine learning methods are known to suffer from the fact that high dimensional spaces are sparsely populated by data points. The number of parameters is varied from 3, 6, 12, 24, to 39, including respectively: $\{c_0, c_1, d_0\}$, $\{c_0 - c_3, d_0, d_1\}$, $\{c_0 - c_5, d_0 - d_3, a_0, a_1\}$, $\{c_0 - c_{11}, d_0 - d_6, a_0 - a_4\}$ and $\{c_0 - c_{12}, d_0 - d_{12}, a_0 - a_{12}\}$, where we tried to mix static and dynamic coefficients.

The second factor is the frame length in the incremental procedure. This can also affect the results as, at each time step, the model is presented with possibly more or less homogeneous data. The frame length was varied from 20 samples (0.2 sec) to 50, 100, 200 and 300 samples.

4. Results

Figure 1 shows the evolution in time of the mixture model for different dimensionalities on the left and for different frame lengths on the right. Table 1 summarises the results for the final models.

As expected the number of clusters increases in time, i.e. when the model is presented with more data. The asymptotic value of the number of clusters depends on the number of variables used. Interestingly, even though this dependency is mostly ascending monotone, the number of clusters obtained with 39 parameters, is lower than with 24. This can be explained noting that the discriminative power added by the last 15 coefficients and contributing to the likelihood of the data given the model is small compared to the negative effect on the Bayes information criterion of adding more model parameters. This effect could be dependent on the amount of data in use.

Table 2: Variation of information of the classifications obtained with different number of coefficients and frame lengths

# coefficients	3	6	12	24	39
3	0	0.358	0.435	0.471	0.488
6		0	0.376	0.428	0.460
12			0	0.366	0.407
24				0	0.320
39	(frame length = 50)				0
frame length	**20**	**50**	**100**	**200**	**300**
20	0	0.215	0.228	0.253	0.252
50		0	0.195	0.241	0.238
100			0	0.236	0.219
200				0	0.222
300	(# coefficients = 39)				0

Regarding the effect of the frame length, the number of clusters increases faster with short frames. This can be considered as a limitation of the procedure, and may be caused by the fact that the number of clusters can only increase, as already mentioned in Sec. 2.1. Another explanation could involve the way, at each time step, the new data is split into subsets depending on how well it is predicted by the current model.

Figure 2 gives an example of the effect of the frame length on the classification. The example contains the phrase "eh, vad" (Swedish for "eh, what"). The segmentations represent the reference phonetic transcription, and the classifications with frame lengths 50, 100 and 300. The thick lines at the bottom represent

the agreement respectively between the pairs of frame length conditions {50,100}, {100,300} and {50,300}, when the randomly assigned class numbers are mapped according to the best match over the whole material. It can be seen that, in spite of the number of clusters being different, there is a good agreement between the different classifications. The effect of frame length needs however to be further investigated.

Finally, as discussed in Sec. 2.2, a measure of consistency between the classifications is given in Table 2, in the form of the variation of information. The high values obtained when changing the number of coefficients (dimensionality) are probably due to the large difference in number of clusters predicted by the different models.

5. Conclusions

This study investigates the behaviour of incremental model based clustering on a set of child-directed speech data. It suggests that the method can be used to simulate the incremental nature of learning, as well as solving the technical problems arising with large data sets. The effects of two factors, namely the dimensionality of the data and the frame length, are also investigated.

The number of clusters predicted by the method increases with the dimensionality up to 24 coefficients. For higher number of dimensions, the number of parameters seems to penalise the introduction of more classes, according to the BIC criterion.

The method predicts higher number of clusters when shorter frames are used. This probably depends on the fact that the number of clusters can only increase for each time step. This could perhaps be avoided if the way the new data at each time step is partitioned into two subsets was adjusted to the frame length.

Finally the agreement between partitions was evaluated with the variation of information method, showing similar distances when the frame length is varied, and distances that increase with the difference in number of clusters when the dimensionality is varied. Classifications with varying frame lengths seem to be in reasonable agreement.

6. Acknowledgements

The work was carried out, at the Centre for Speech Technology, within the MILLE project funded by The Bank of Sweden Tercentenary Foundation K2003-0867.

7. References

[1] F. Lacerda, U. Sundberg, R. Carlson, and L. Holt, "Modelling interactive language learning: Project presentation," in *FONETIK*, 2004, pp. 60–63.

[2] F. Lacerda, E. Klintfors, L. Gustavsson, L. Lagerkvist, E. Marklund, and U. Sundberg, "Ecological theory of language acquisition," in *EPIROB*, 2004, pp. 147–148.

[3] S. Deligne and F. Bimbot, "Inference of variable-length acoustic units for continuous speech recognition," in *ICASSP*, vol. 3, 1997, pp. 1731–1734.

[4] T. Holter and T. Svendsen, "Combined optimisation of baseforms and model parameters in speech recognition based on acoustic subword units," in *ASRU*, 1997, pp. 199–206.

[5] S. Watanabe, A. Sako, and A. Nakamura, "Automatic determination of acoustic model topology using variational bayesian estimation and clustering," in *ICASSP*, vol. 1, 2004, pp. 813–816.

[6] C. Li and G. Biswas, "Temporal pattern generation using hidden markov model based unsupervised classification," in *Advances in Intelligent Data Analysis: Third International Symposium*, vol. 1642, 1999, p. 245.

[7] T. Oates, L. Firoiu, and P. R. Cohen, "Clustering time series with hidden markov models and dynamic time warping," in *IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning*, 1999, pp. 17–21.

[8] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Application of variational bayesian approach to speech recognition," in *Advances in Neural Information Processing Systems 15*, S. T. S. Becker and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 1237–1244.

[9] F. M. Porikli, "Clustering variable length sequences by eigenvector decomposition using HMM," in *International Workshop on Structural and Syntactic Patt. Rec.*, 2004.

[10] N. Cowan, "On short and long auditory stores," *Psychological Bulletin*, vol. 96, no. 2, pp. 341–370, 1984.

[11] G. McLachlan and K. Basford, *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1988.

[12] J. D. Banfield and A. E. Raftery, "Model-based gaussian and non-gaussian clustering," *Biometrics*, vol. 49, no. 3, pp. 823–835, 1993.

[13] C. Fraley and A. E. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis," *Computer Journal*, vol. 41, no. 8, 1998.

[14] A. Dasgupta and A. Raftery, "Detecting features in spatial point processes with cluster via model-based clustering," *J. Amer. Statist. Assoc.*, vol. 93, no. 441, pp. 294–302, 1998.

[15] C. Fraley, "Algorithms for model-based gaussian hierarchical clustering," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 270–281, 1998.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Stat. Soc. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[17] C. Fraley, A. Raftery, and R. Wehrensy, "Incremental model-based clustering for large datasets with small clusters," Department of Statistics, University of Washington, Tech. Rep. 439, 2003.

[18] R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[19] C. Fraley and A. Raftery, "MCLUST: Software for model-based clustering, density estimation and discriminant analysis," Department of Statistics, University of Washington, Tech. Rep. 439, 2003.

[20] M. Meilă, "Comparing clusterings," Department of Statistics, University of Washington, Tech. Rep. 418, 2002.

[21] F. Lacerda, E. Marklund, L. Lagerkvist, L. Gustavsson, E. Klintfors, and U. Sundberg, "On the linguistic implications of context-bound adult-infant interactions," in *EPIROB*, 2004, pp. 149–150.

[22] L. Gustavsson, U. Sundberg, E. Klintfors, E. Marklund, L. Lagerkvist, and F. Lacerda, "Integration of audio-visual information in 8-months-old infants," in *EPIROB*, 2004, pp. 143–144.