55

UNSUPERVISED LEARNING OF SIMPLE SPEECH PRODUCTION BASED ON SOFT COMPETITIVE LEARNING

Georg Dorffner, Thomas Schönauer

Dept. of Medical Cybernetics and Artificial Intelligence
University of Vienna, Freyung 6/2, A-1010 Vienna, Austria
and Austrian Research Institute for Artificial Intelligence
georg@ai.univie.ac.at

55.1 INTRODUCTION

In this paper we present a simple connectionist model for the adaptive sensory-motor loop involved in perceiving and producing speech. At the heart of the production part lies an articulatory model which approximates the human vocal tract through polygons and splines. Output of this model is the envelope of the acoustic filter function, realized by this vocal tract, which is comparable to the spectrum of real speech segments. The goal of this research was to find a learning method to train a multi-layer neural network to produce the correct set of twelve articulatory parameters when given the spectrum of recorded real speech (stationary vowels). The method introduced in this paper explicitly makes use of a neural network categorization component. Through so-called *soft competitive learning* it learns to gradually compress the responses to more and more unitized categorical patterns. After a precategorization phase, during which presented real speech patterns are classified, the model starts to randomly produce output signals. A goodness-of-fit measure, which can be computed easily, is taken as the criterion whether the self-produced signal is close enough to any of the known categories, and as the learning rate to adapt the weights between the categorization layer and the output units.

55.2 THE ARTICULATORY MODEL

The used articulatory model implements a geometrical approximation of the midsagittal image of the human vocal tract (Fig. 55.1). The tongue body is approximated by a circle with variable center and radius. The other parts are

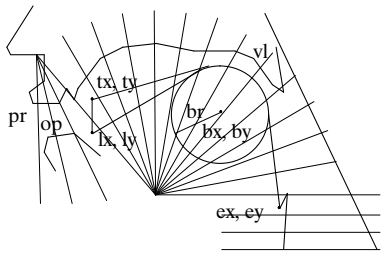

Figure 55.1 The articulatory model

seen as straight lines or polygons. Each configuration of the vocal tract can be uniquely described with a set of twelve parameters. A grid of 25 intersecting lines is used to compute the distances between upper and lower articulators. From this distance profile the cross-sectional areas are computed. Then the vocal tract is viewed as consisting of 24 tube segments which together behave like a filter. By computing the reflection coefficients at each border between two segments (based on the areas), and LPC coefficents, one can finally arrive at a set of frequency values representing the envelope of this filter. These values can be taken as corresponding to the speech signal produced by this model.

55.3 THE NEURAL NETWORK MODEL

The problem of teaching a network to produce the parameters of the articulatory model is an instance of the classical reinforcement learning problem in motor control (see e.g. [5]). The idea behind our solution is based on the follwing observations.

- Learning should mainly be based on receiving external examples and on monitoring the system's own performance (self-supervised learning).
- The criterion for evaluating the error in order to make the system adapt should not be what is similar in the physical world, but instead, what the system *perceives* as similar.
- The acquisition of sounds should be guided by recognizing that they can be divided into several categories (the phonemes).

The chosen connectionist approach is depicted in fig. 55.2. The core part of the

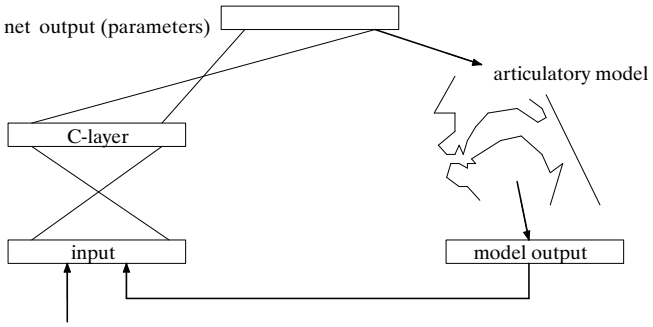

Figure 55.2 The overall connectionist architecture

network is a layer that performs *soft competitive learning* to categorize stationary speech signals in the frequency space presented at the input layer (25 units) into distinct categories. The output layer consists of 12 units corresponding to the input parameters of the articulatory model (appropriately scaled into values in the range [0..1]). The next two subsections describe how learning is done.

55.3.1 Soft competitive learning

What distinguishes *soft* from regular competitive learning (such as [6, 2]) is that the responses in this layer (called *C-layer*) initially are distributed. For this a C-layer has full intralayer connections and operates with the interactive activation rule ([4]). These connections are initialized with zero or small negative (inhibitory) weights. This achieves the well-known *rich-get-richer* effect between units in this layer, which tends to strengthen highly active units and weaken others. As in hard competitive learning, the winning (i.e. the most highly active) unit is chosen and its connections are strengthened. In addition, however, the other connections are also adapted. In mathematical terms, the learning rules are implemented the following way.

$$\Delta w_{ij} = \begin{cases} \eta(x_i - w_{ij}) & \text{if unit } i \text{ is the winner} \\ \eta(\mu_j x_i + (\lambda_j - 1) w_{ij}) & \text{otherwise} \end{cases}$$

$$\mu_j = -\frac{net_j}{\sqrt{1-net_j^2}} \quad \lambda_j = \frac{1}{\sqrt{1-net_j^2}}$$

where net_j is the net input (the weighted sum) of the j-th unit in the C-layer. The weight vector of the winner is moved toward the input vector (see [2]). The two factors μ and λ are derived from the assumption that the non-winning weight vectors should approach a vector orthogonal to the input. Both weight and input vectors are normalized to length 1. Finally, intra-layer connections increase their inhibitive effects through inverse Hebbian learning. This assures that the final responses (after competition) are compressed.

55.3.2 Goodness-of-fit

The above learning rule assures that only input vectors close enough to the learned prototype lead to a quasi-unitized response. Thus by quantifying how unitized a pattern is we can easily derive a goodness-of-fit value g^p indicating how well a given input pattern p belongs to the class.

$$g^p = \frac{1}{2}((x_w^p - \frac{1}{n-1}\sum_{i \neq w} x_i^p) + (x_w^p - x_r^p))$$

where x_w is the activation of the winner, x_r that of the runner-up (the second-highest activation); n is the number of units in the C-layer. The first part in this formula reflects the difference between the winner and the average of the other activations in the C-layer. The second part ensures that indeed only a single highly active unit leads to a large goodness-of-fit. Whenever g^p reaches a value above a threshold θ the weights between the winner of the C-layer and the output layer (denoted by v) are adapted according to the outstar rule ([3]), using the difference between g^p and θ as an additional learning rate.

$$\Delta v_{wj} = \eta(g^p - \theta)(x_j - v_{wj}) \quad \text{for all } j$$

The value of g^p also controls the random component producing the articulatory parameters during the exploratory learning phase. After each successfully categorized sound a few cycles are added where Gaussian noise is added to the previous set of parameters, with a standard deviation indirectly proportional

to g. The function of this learning scheme can be described as follows. First randomly, then more and more guided by previously recognized phonemes, the system continually produces speech signals. These signals are categorized by the C-layer, which was trained on real speech. Whenever it hits one of the categories, expressed by a large g, the weights between the winner and the output are adapted, gradually associating it with the set of parameters which produced the signal.

55.4 RESULTS

During categorization of external input the network was trained with 4 recorded and Fourier transformed instances of each of 5 distinct German vowels. Depending on the number of units in the C-layer and the weight initialization between 3 and 15 classes were learned. Of course, ideally five classes should be learned. But as it turned out, these classes cannot be defined naturally through Euclidean distance, pointing to the need of more complex preprocessing of the speech data. For the sake of the experiments this was of no importance. In the subsequent phase of exploratory self-supervised learning between 100 and 500 cycles were performed in each training task. The random generator was controled as described above. After each trial with one random sound, five more trials in the "neighborhood" (defined through the Gaussian noise) were performed in each cycle. The parameters varied between learning tasks were the threshold θ and the number of pre-learned classes.

The network successfully learned to reproduce the prototypes of between 20 and 70 % of the classes it had recognized in the first phase. The higher θ the closer the reproduced signals were, due to the fact that only very good fits lead to adaptation. Also, the results were better when the number of pre-learned classes was large, although this came with a lower overall percentage of learned sounds. Two main reasons can be identified why the results were not better in these experiments. First, the articulatory model appears to be still incapable of producing some of the desired sounds. Secondly, the problem of non-uniqueness (two different sets of parameters leading to the same signal) cannot be solved in this simple network, leading to bad solutions where this is the case. Nevertheless, the results demonstrate the validity of the core approach of reinforcement learning based on soft categorization.

55.5 DISCUSSION

Among others, the approach is interesting for the following reasons. First, the

criterion for goodness-of-fit is moved *inside* the model. In other words, the network learns whenever it recognizes something as similar, and not when a subjective distance measure in the physical signal is low. Secondly, the model explains some psychological phenomena, such as the loss of sensitivity toward subtle variations in sounds, once categorization has gone beyond a certain level. In [1] a similar approach is described. There, also the importance of partially compressed responses is stressed. Our approach differs from theirs in that it is simpler and achieves a goodness-of-fit in a strictly feedforward manner, bypassing the feedback necessary for resonance in ART. This might have negative implications what stability of the categories is concerned, which however was not our major concern in this work.

55.6 ACKNOWLEDGMENTS

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Science and Research.

REFERENCES

[1] Cohen M.A., Grossberg S., Stork D.G.: Speech Perception and Production by a Self-organizing Neural Network, in Lee Y.C.(ed.), *Evolution, Learning and Cognition*, World Scientific Publishing Co., London, p. 217-231, 1989.

[2] Grossberg S.: Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors, *Biological Cybernetics* 21, 145-159, 1976.

[3] Grossberg S.: *Studies of mind and brain*, Reidel Press, Boston, 1982.

[4] McClelland J.L., Rumelhart D.E.: An Interactive Activation Model of Context Effects in Letter Perception: Part 1. An Account of Basic Findings, *Psychological Review* 88, 375-407, 1981.

[5] Miller T.W., Sutton R.S., Werbos P.J.(eds.): *Neural Networks for Control*, MIT Press, Cambridge, 1991.

[6] Rumelhart D.E., Zipser D.E.: Feature Discovery by Competitive Learning, *Cognitive Science* 9(1)75-112, 1985.