MULTI-SPEAKER VOWEL CLASSIFICATION WITH ADAPTIVE NEURAL NETS

David Weenink and Louis Pols
Institute of Phonetic Sciences, University of Amsterdam, The Netherlands

ABSTRACT

We studied vowel classification and speaker normalization performance with neural nets based on Adaptive Resonance Theory (ART). ART was developed by S. Grossberg [4] as a theory of human cognitive information processing. It is the result of an attempt to understand how biological systems are capable of retaining plasticity throughout life, without compromising the stability of previously learned patterns. We have implemented some of these ideas in a supervised neural network called CategoryART [9]. The neural network was trained with formant frequency values extracted at the midpoint of vowels. Vowels were selected from the TIMIT speech corpus [6]. Separate train and test sets were used. Of the 630 speakers in this database 438 are male and 192 are female. A simple preprocessing algorithm achieved normalization. Only the 13 monophthong vowel categories (iy, ih, ey, eh, ae, aa, ow, ah, ao, uw, uh, ux, er) were used. Formant frequency values were determined by an LPC analysis. To compare formant frequency values for males and females, normalized frequency values were calculated in a preprocessing stage. Next to the neural net we also used a Gaussian classifier. This classifier attained on the average 57% correct classification. The neural network did not perform as well as the Gaussian classifier and only achieved 50% correct classification.

1. INTRODUCTION

For spoken language understanding, phonetic classification is probably one of the fundamental abilities needed. A very promising theory about human cognitive information processing is Adaptive Resonance Theory (ART) which was developed by Grossberg [4]. Basic features of Adaptive Resonance Theory and its relation to perception are laid out in a great number of articles by Grossberg and his associates (see for example Grossberg 1986 for an overview). ART was the result of an attempt to understand how biological systems are capable of retaining plasticity throughout life, without compromising the stability of previously learned patterns. Somehow biologically based learning mechanisms must be able to guard stored memories against transient changes, while retaining plasticity to learn novel events in the environment. This tradeoff between continued learning and buffering of old memories has been called by Grossberg the stability-plasticity dilemma. This poses special design problems, since, for example, in (supervised) feedforward networks, which are the most popular neural networks nowadays, new information gradually washes away old information, and therefore, feedforward networks cannot be made stable in a changing environment.

To be able to mimic biological behaviour, the emphasis of ART neural networks lies at unsupervised learning and self-organization. Unsupervised learning means that the network learns the significant patterns on the basis of the inputs only, there is no feedback. There is no external teacher that instructs the network to which category certain input belongs. Other types of learning are reinforcement learning and supervised learning. In reinforcement learning the net receives only limited feedback, like "(on this input) you performed well" or "(on this input) you have made an error". In supervised mode a net receives for each input the correct response. Unsupervised learning is the substrate on which the other types of learning are based. Learning in biological systems always starts as unsupervised learning: for the newly born hardly any pre-existing categories exist. A system that can learn in unsupervised mode can always be adjusted to learn in the other modes, like reinforcement mode or supervised mode. However, a system specifically designed to learn in supervised mode can never perform in unsupervised mode. Needless to say that in unsupervised mode we cannot have a separate training and performance phase because this implies the presence of a homunculus that knows when to alter phases. Self-organization means that the system must be able to build stable recognition categories in real-time. These design constraints have led to a series of real-time ART neural network building blocks for unsupervised category learning and pattern recognition.

By providing an unsupervised ART neural network building block with external feedback, we can make it reorganize its recognition categories and in this way incorporate supervision. In the ARTMAP architecture [2], two unsupervised building blocks, together can implement supervision by letting one module give feedback to the other. We have made a simple algorithmic version of such a supervised neural network and named it CategoryART [9]. The feedback mechanism is called match tracking. Just like ARTMAP, CategoryART was designed to conjointly maximize predictive success and minimize predictive error by linking predictive success to learned category size. In here, we will test some of the strengths and weaknesses of CategoryART on a classification test with vowels from the TIMIT database. More specifically, we will compare the phonetic vowel classification performance of CategoryART, in the absence of word recognition and based on local acoustic information, with classification from a Gaussian classifier.

2. THE CATEGORYART NEURAL NETWORK

The CategoryART neural network is a supervised two-layer neural network. It incorporates a FuzzyART building block [3] and uses external feedback. In this building block all nodes in the first layer are connected by bottom up and top down paths with all the nodes in the second layer. The paths have associated connection strengths called weights that represent the long-term memory of the system. Initially all connections strengths have some small random value. As learning proceeds, some of the nodes in the second layer may become committed and their weights represent prototype information. Committed nodes are labelled with one of the labels in a labelSet. Initially the labelSet

is empty. When a labelled input pattern I is presented to the network and its label is not yet known, a new label is added to the labelSet and a new uncommitted node learns the pattern I. When the label is known the input I is matched with all the learned prototypes and when the best match is "good enough" this prototype is a modified a little bit in the direction of pattern I. When the match is not good enough a new node is committed that learns the pattern I. The "good enough" criterion is controlled by a dimensionless parameter ρ called vigilance. The modification of the prototype is controlled by a learning parameter β. It is important to note that learning occurs when there is a *match* between the input pattern and a prototype and not on the basis of a mismatch. The learning algorithm in pseudo-code goes as follows:

```
for numberOfEpochs
    for all (pattern I, label c) in data
        learn (I, c)
    endfor
endfor

procedure learn (pattern I, label c)
    if label c not in labelSet
        add label c to labelSet
    endif
    J = find_winning_node (I)
    if labelSet [node[J]] _ c && matchtrack
        temporarily increase vigilance
        J' = find_winning_node (I)
        reset vigilance
    endif
    update_weights w[J']
    node [J'] = index label in labelSet
endprocedure
```

Weights are updated according to the following equation

$$\mathbf{w}_J^{(new)} = \beta(\mathbf{I} \wedge \mathbf{w}_J^{(old)}) + (1-\beta)\mathbf{w}_J^{(old)},$$

where the subscripts *old* and *new* refer to the weights before and after updating. **I** is the input pattern and β the learning parameter. When β=1 we speak of fast or one-shot learning. When fast learning is enabled the weight vector **w** for the newly committed node equals the input pattern. After the commitment the weight vector update causes the new weight vector to become more aligned with the most recently coded input pattern.

3. DATA SELECTION

The TIMIT acoustic phonetic speech corpus contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States of America, 438 speakers were male, 192 were female. The corpus is divided in a *train* and a *test* part. All the sound files in the TIMIT database have an accompanying label file that contains start and end sample numbers of all phonemes occurring in the sentence. We used these label files to construct a phoneme database with an entry for every single phoneme label in TIMIT. This resulted in a database with 241225 entries. Each entry contains information about, among others, the dialect, the speaker, the sentence, the left and right context, the stress value and the begin and end time with respect to the beginning of the sentence [8]. This database enabled us to calculate the midpoint of every individual vowel. From the 20 different vowels present in the database we selected the same 13 monophthong vowels in American English as did Meng and Zue [7]. These vowels have TIMIT labels iy, ih, eh, ey, ae, aa, ah, ao, ow, uh, uw, ux, er. These translate into the IPA labels i, ɪ, ɛ, e, æ, ɑ, ʌ, ɔ, o, ʊ, u, ü, ɜ˞, respectively. The separation of TIMIT into a train and test part, and the male-female marking of the files in this database, naturally leads to 4 different vowel sets: male speakers in the train part (mtrain: 26338 vowels), and the test part (mtest: 9047 vowels), female speakers in the train part (ftrain: 11288 vowels) and female speakers in the test part (ftest: 4741 vowels). Although the distribution of the number of occurrences of each phoneme doesn't differ much between these 4 sets, the number of occurrences of each phoneme within each set differs widely. However, our classifying methods can be made robust against these differences.

The first three formant frequency values were measured for all vowels in these 4 vowel sets at the vowel midpoints in an automatic way. The following measuring procedure was performed on all sound files with the speech analysis program Praat [1]:

- Downsampling. Sound files from female speakers were downsampled from 16 kHz to 11 kHz and files from male speakers were downsampled to 10 kHz. This reduces the maximum number of formants in the signal to approximately 5.
- Pre-emphasis, followed by LPC analysis based on autocorrelation, with 25 ms window length, 10 filter coefficients and 5 ms time step.
- Solving for the maximally 5 formant candidates in each analysis frame.
- Query the formant tracks for the 3 (lowest) frequency values at the midponts of the vowels by means of linear interpolation. These 3 frequency values we associated with the first 3 formant frequencies.

For each of the 4 vowel sets defined above, vowel identity and the three formant frequency values were collected row-wise into a table. In this way the i-th column of a table contains all the i-th formant frequency values. In order to improve on the distributional properties of the formant frequencies, for all tables the logarithm of the formant frequencies were calculated. This transformation brings the "variance" of all formant frequencies approximately into the same range. Next we centre the tables by column: for each column we first calculate the average value of the log(F)'s and subtract this value from all entries in the corresponding column. As a result we know that for the 4 datasets the average value for all formants equals zero. This has a nice normalization effect because it brings formant frequency values from male and female speakers into the same range as can clearly be seen from fig. 1b. As a consequence, all formant frequency values for the first three formants now lie approximately in the non-symmetric interval [-0.41, +0.28]. 82 vowels whose formant frequency value were outside this interval, mainly male i, u and ü whose first formant could not be measured, were left out.

Because the input values for the neural net must all lie in the interval [0, 1] a last linear transformation was performed by first adding 0.41 to all entries in the tables and subsequently dividing by 0.69 (=0.41+0.28).

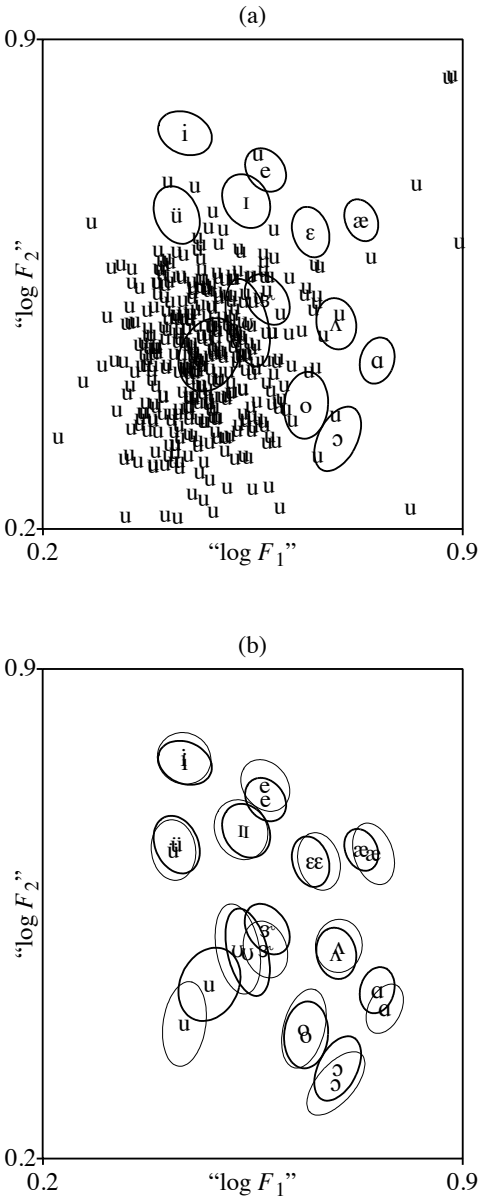

Figure 1. (a) 1/2-σ ellipses in a plane spanned by the first two formant frequencies. The ellipses were determined for 13 monophthong American English vowels, based on a selection with 26308 entries from 326 male speakers as calculated from the TIMIT train database. The real distribution of the 396 /u/'s in this selection is also shown. The logarithmic horizontal and vertical frequency axes were scaled such that all frequencies are within the [0, 1] interval. (b) The same 1/2-σ ellipses as above (fat contours) together with the ellipses as determined for the 136 female speakers in the train database after normalization (11288 entries).

4. TESTING PROCEDURE

For the two training tables a discriminant analysis was performed with the Praat program. We then obtain for each vowel a sscp matrix with sums of squares and cross products. These 13 individual sscp matrices were pooled and we obtained the "pooled sscp matrix" whose inverse is used in classification with the Mahalanobis distance measure. From the number of occurrences of each phoneme in the train set, a priori probabilities were calculated. The Gaussian classifier uses both the inverse of the "pooled sscp matrix" as well as the a priori probabilities in the distance calculation.

The training procedure for the neural nets is as follows. Learning consists of presenting each pattern accompanied by its category label a number of times to the neural network. Each presentation of the whole training set is called an epoch. We always used 20 epochs, which means that each pattern was presented exactly 20 times to the network. After each epoch a new permutation of the training set was presented to the network. In ART theory there is no principal difference between a network in the learning phase and in the classification phase. However, for testing purposes we made a switch to be able to separate these two phases. In these experiments we kept learning rate $\beta = 0.1$ and vigilance $\rho = 0.9$. Because of the combination of match tracking and fast learning, a single CategoryART system can learn a prediction for a rare event that is different from that for a cloud of similar frequent events in which it is embedded. This possibility is too strong given the spread in the data and generates too many scattered prototypes. This would result in poor generalization, so we learn with match tracking off.

5. RESULTS

To gain some insight in the amount of variation of vowel categories with formant frequency values from different speakers, we calculated for each vowel category the covariance matrix from the normalized log(F) values. These covariance matrices are necessary to calculate the σ ellipses. A 1-σ ellipse theoretically covers 68.3% of the data, provided the data were drawn from a multi-normal distribution. In figure 1a are shown the 1/2-σ ellipses for the vowels from the male speakers in the train part of the dataset (mtrain). These ellipses cover approximately 38.3% of the data. To give some insight in the real distribution of the vowels, we have also plotted all the 396 /u/'s that are present in this set. The amount of spread is considerable and we can clearly not hope for perfect recognition scores. The /u/'s at the upper right have formant frequency misfits because their first formant could not reliably be determined. Fig 1b shows the 1/2-σ ellipses for the vowels in the male (fat contours) and female train datasets (mtrain and ftrain). We note the remarkable coincidence in distributional properties for the male and female data. Almost perfect overlap of the ellipses from the same vowel category. The between-speaker-category variance almost disappeared we are only left with still considerable within-speaker-category variance.

The percentages correct of the classification experiments are shown in table 1, row labels shows the dataset used for training and column labels shows the datasets used for classification. The upper two rows in the table show results obtained with the Gaussian classifier, the lower two rows the results with the neural network model.

Table 1. Percentage correct scores for Gaussian classifier (upper 2 rows) and CategoryART neural net classifier (lower 2 rows).

	mtrain	mtest	ftrain	ftest
mtrain	55.1 ()	57.3 ()	53.0 ()	52.6 ()
ftrain	54.5	56.5	54.8	54.5
mtrain	51.0	48.3	52.3	49.9
ftrain	48.8	49.9	52.4	49.9

5. DISCUSSION

There are several reasons why the classification results from the CategoryART neural network fall somewhat behind the results obtained with the Gaussian classifier. As can be seen from figures 1a and 1b the overlap between different vowel categories is considerable. A Gaussian classifier uses all the data at the same time to make a suitable separation of the space in contiguous areas. A neural net classifier only uses one data point at a time. A second possible cause for the weaker performance of CategoryART is its implementation of one-shot learning. It is now obvious that simply using one-shot learning is not optimal when the data contains (much) variance (noise). The one-shot learning option makes the network sensitive to noise: actually noise will be learned.

In our future research we will try out a number of possible improvements on the neural net part and on the data analysis part. First of all, to improve on the dynamics of CategoryART, we could actually turn the one-shot learning to our advantage by a procedure called majority-voting [2]. We train a number of networks, say 5 with the same data but with a different random ordering of the individual patterns. Then we let them classify some data. The label that each network assigns can be considered as a vote. The label with the largest number of votes wins. As, among others, Meng and Zue [7] achieve better classification results by incorporating dynamic information this will be incorporated as well. Next we have to further improve on formant frequency measurements by imposing local continuity constraints on formant frequency values distilled from LPC analyses with orders higher than 10. A normalization model in which a better modelling of the similarities and differences of speaker characteristics is possible could of course obtain further improvements.

REFERENCES

[1] Boersma, P.P.G. and Weenink D.J.M. (1986), *Praat: A system for doing phonetics by computer*, Report of the Institute of Phonetic sciences of the University of Amsterdam 132, (http://www.fon.hum.uva.nl/praat).

[2] Carpenter, G.A., Grossberg, and Reynolds, J.H. (1991), ARTMAP: Supervised real-time learning and classification of non-stationary data by a self-organizing neural network, *Neural Networks* 4, 656-588.

[3] Carpenter, G.A., Grossberg, S. and Rosen, D.B. (1991), Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system, *Neural Networks* 4, 759-771.

[4] Grossberg S. (1976), Adaptive pattern classification and universal recoding I: Parallel development and coding of neural feature detectors, *Biological Cybernetics* 23, 121-134.

[5] Grossberg, S. (1986), The adaptive self-organization of serial order in behavior: speech, language and motor control, in *Pattern Recontion By Humans And Machines*, Volume I: Speech Perception, E. Schwab & H. Nusbaum (eds.), Academic Press, Inc.

[6] Lamel L.F., Kassel, R.H. and Seneff, S. (1986), Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus, *Proc. DARPA Speech Recognition Workshop*, Report NO. SAIC-86/1546, 100-109.

[7] Meng, H.M. and Zue, V.W. (1991), Signal representation comparison for phonetic classification, IEEE Proceedings of ICASSP 91, Toronto, 285-288.

[8] Weenink, D.J.M. (1996), Adaptive vowel normalization and the TIMIT acoustic phonetic speech corpus, *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 20, 97-110.

[9] Weenink D.J.M. (1997), Category Art: A variation on Adaptive Resonance Theory Neural Networks, *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21, 117-129.