Vowel Recognition and (Adaptive) Speaker Normalization

Louis C.W. Pols and David J.M. Weenink

Institute of Phonetic Sciences / ACLC
Herengracht 338, 1016 CG Amsterdam, The Netherlands

{Louis.Pols; David.Weenink}@uva.nl

Abstract

In automatic and human speech recognition alike, there is the unsolved problem of non-uniqueness in speech production because of many sources of more or less systematic variability (global and local context, speakers, style, communication channel, etc.) *versus* the supposed fixed distributional variance in template-based recognition. This paper concentrates on acoustic vowel recognition (neglecting language modeling) and various ways of extrinsic and intrinsic adaptive speaker normalization. We used such methods as discriminant analysis, Procrustes transform, feedforward neural nets and adaptive resonance theory networks. One of the problems is to optimize fast learning, adapting and generalizing on the basis of small amounts of new information, *versus* not forgetting acquired knowledge too quickly. For training and testing we used Dutch databases of vowels spoken by men, women and children, plus the vowel segments from the American-English TIMIT database with 438 male and 192 female speakers.

1. Introduction

One of the amazing capabilities of human speech recognition and understanding is its robustness to all sorts of variability. This paper will actually concentrate on the rather limited but still challenging topic of *vowel* recognition. Even for these, generally high-intensity and all-voiced, speech segments the robustness problem holds equally well. Imagine the variability caused by different speakers with markedly different vocal-tract and vocal-cords sizes, the contextual variability, effects of age, speaking rate and speaking style, influence of the communication channel upon the spectro-temporal characteristics of vowels, etc. Most emphasis in this paper will be given to *between-speaker variability*. A further limitation in this paper is our restriction to high-quality recordings of read speech. This still leaves as major sources of variability: speaker category (male, female, child) and individual speakers, context, stress, speaking rate and speaking style. Also the order in which subsequent stimuli are being presented (blocked per speaker or speakers mixed) for identification appears to be important.

After having reviewed human vowel recognition in sect. 2, we will study in sect. 4 how well automatic vowel recognition does perform if only acoustic information is available. This can be so-called *stationary* (single frame) or *dynamic* (multi-frame) information. Our limitation to *acoustic* information only indicates that language modeling plays no role in our present approach. In a later phase such information can of course always be added to improve overall performance.

Unavoidably our vowel recognition scores will thus be rather low. However, the *relative improvement* from one condition, or one approach, to another will be more important to us.

In sect. 3 we will first introduce various recognition approaches, such as discriminant analysis, Procrustes transform, feedforward neural nets and adaptive resonance theory networks. Contrary to most present-day speech recognition systems, we did not include HMM-based single-phone or multi-phone models. Furthermore we will indicate how extrinsic or intrinsic speaker normalization can be incorporated in these techniques. *Extrinsic* normalization is considered to be based on preliminary information about the speaker, such as speaker-specific average vowel positions. *Intrinsic* normalization is based exclusively on vowel and speaker information of the momentary speech segment.

2. Human vowel recognition

A native speaker of the language will generally only make vowel recognition errors under critical conditions. This may have to do with sloppy pronunciation, poor signal conditions, totally unexpected (change in) topic of conversation, unknown words, etc. In the context of the present paper we will concentrate on one other possible source of poorer performance, namely the blocked *versus* mixed presentation of a series of vowel stimuli from various speakers. We recorded 10 male and 10 female speakers, as well as 10 children while pronouncing the Dutch utterance '*V van pVt*', with all 12 Dutch monophthongs as V. Both V-segments per utterance were isolated and then presented under various conditions to 20 listeners for identification. Actually in the listening experiments only half of the speakers (3 x 5) were used, for more details, see [12]. In the *blocked* conditions all vowel stimuli from one speaker were presented sequentially, whereas in the *mixed* conditions each subsequent vowel stimulus was from a different speaker.

Table 1 summarizes the error scores, averaged over 20 subjects, 12 vowels and 15 speakers, while distinguishing between mixed (M) and blocked (B) condition, for each of 8 different experiments (for more details, see [12]). Most stimuli had a fixed duration of 50 ms, which precludes a proper distinction between long and short vowel pairs. Each short vowel response given to its long counterpart stimulus (such as /O/ for /o/) was thus considered to be a correct response, however, when the reverse happened this was still considered to be an error. The overwhelming conclusion of these results can only be that under the blocked condition listeners always performed better than under the mixed condition. For other interesting observations we refer to [12,13].

Table 1: Error percentages for vowel identification, averaged over subjects (20), vowels (12) and speakers (15 in total, or 5 per group men, women or children). Results are split for mixed (M) and blocked (B) presentation of the stimuli. The scores have been corrected for long/short confusions. For more details, see text.

Expt.	Description	Averaged		Men		Women		Children	
		M	B	M	B	M	B	M	B
1	full V	9.6	3.8	8.5	3.8	10.6	3.6	9.8	4.0
2	50 ms from V	18.7	15.1	12.9	11.1	16.1	12.7	27.1	21.6
3	50 ms from pVt	24.2	18.1	22.0	15.1	21.0	17.7	29.6	21.5
4	50 ms, mean F0	29.0	25.8	22.6	20.8	24.9	20.7	39.3	35.9
5	50 ms, F0=135 Hz	36.7	28.2	25.0	17.3	28.0	23.3	57.0	44.0
6	50 ms, F0=235 Hz	36.7	29.0	36.3	28.0	26.1	22.0	47.9	37.0
7	50 ms, F0=335 Hz	49.2	45.8	62.8	59.3	42.5	40.4	42.4	37.8
8	50 ms, noise	34.5	26.0	30.5	19.6	25.3	21.2	47.8	37.4

3. Methods for vowel recognition and for adaptive speaker normalization

Most ASR-people will immediately consider Markov models (HMM) and/or neural nets (ANN) to be the best or perhaps even the only possible approach for phoneme recognition. However, we were not interested in the integration of such an acoustic recognizer into a full-fledged ASR-system. On the other hand we *did* want to have full control over number of vowel categories, type of parametric representation, dimensionality, type of training and speaker adaptation, etc. So, we tested and further developed a number of other pattern recognition procedures, that we will summarize below. We will also summarize the two parametric representations that we mainly used, namely formant analysis and bandfilter analysis followed by some form of data reduction. For many more details, see [13].

3.1. Parametric representation of spectral information

The most popular vowel representation in Phonetics is a *formant* representation [1]. Undoubtedly a 3- to 5-formant vowel representation is highly informative, quite unique and highly noise-resistant. However, a fully automatic and error-free formant analysis is still not available. Furthermore, it is much more a representation of the resonance characteristics of the speech *production* process, rather than a proper reflection of the peripheral processing in the inner ear. That is probably better achieved by a *bandfilter* representation. Such a representation is of a much higher dimensionality, that however can be reduced by applying for instance a principal components analysis (maximizing the amount of variance explained), or a discriminant analysis (maximizing class distinctions) or MFCC (a fourier-like decomposition of the envelope spectrum). As far as the available databases permitted, we tested both a formant representation as well a bandfilter representation of our vowel segments. Most learning models that we implemented (see sect. 3.3) have no explicit notion of time, but they can handle spectral vowel data very well, although they do not explicitly model vowel dynamics. However, we sometimes simulated dynamicity by grouping several analysis frames into one data item.

3.2. Dutch vowel databases and American-English TIMIT

In sect. 2. we already gave some information about the small size Dutch vowel database. It is minute compared to TIMIT, but its structure is optimally tuned to the needs for this project: it contains vowel segments from three highly different speaker groups (10 men, 10 women and 10 children) under two well controlled conditions (vowels in isolation, and vowels in pVt context). Furthermore, hand-edited formant measurements are available (for average F1-F2 positions, see Fig. 1) and bandfilter analyses can easily be achieved via *praat* [2].

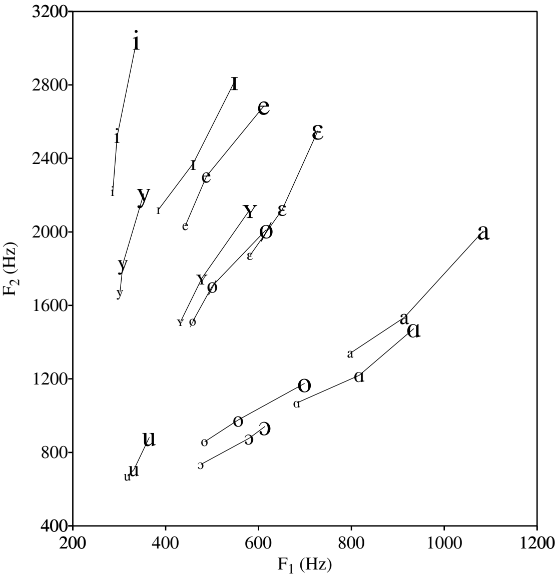

Figure 1: Average F1-F2 vowel positions for the 10 male (small), 10 female (medium) and 10 children (large size symbols) Dutch speakers.

Another Dutch vowel database that we used concerns 50 male [9] and 25 female speakers [10] pronouncing Dutch vowels in hVt. Unfortunately the original audio recordings are no longer available, so we can only rely on the freely available formant measurements.

Finally we extensively used the freely available TIMIT database. The audio recordings of 630 American English speakers, each pronouncing 10 sentences, are extended with a (hand-made) transcription and segmentation at the sentence, word and phoneme level. The stressed and unstressed vowel segments together comprise 78,374 vowel segments in 20 vowel categories, the subset of *stressed* vowels contains 49,562 segments. For some more details, see Table 2. The second author made this database accessible in *praat* [2].

Table 2: Number of speakers and of vowels in the TIMIT database, split up for male and female speakers and for train and test set. The so-called summary data contain up to 20 average vowels per speaker. For more details, see text.

		Male	Female	Total
	nr. speakers	*326*	*136*	*438*
train	nr. vowels	40,468	16,995	57,463
	summary	6,008	2,490	8,498
	all stressed	25,706	10,622	36,328
	nr. speakers	*112*	*56*	*192*
test	nr. vowels	13,889	7,022	20,911
	summary	2,070	1,011	3,081
	all stressed	8,845	4,389	13,234
	nr. speakers	*438*	*192*	*630*
total	nr. vowels	54,357	24,017	78,374
	summary	8,078	3,501	11,579
	all stressed	34,551	15,011	49,562

3.3. Various pattern recognition procedures

Most of the pattern recognition procedures that we applied are not new, but some of them are not generally considered to be recognition procedures, but only dimensionality reduction procedures. For more details than can be given in this conference contribution, we refer to [13]. It should also be mentioned that these procedures plus several data sets are all available within *praat*.

3.3.1. Principal components analysis (PCA)

PCA learns from *unlabeled* multi-variate data how to represent these data in an efficient way, such that subsequent new dimensions are found in which the variance is maximal.

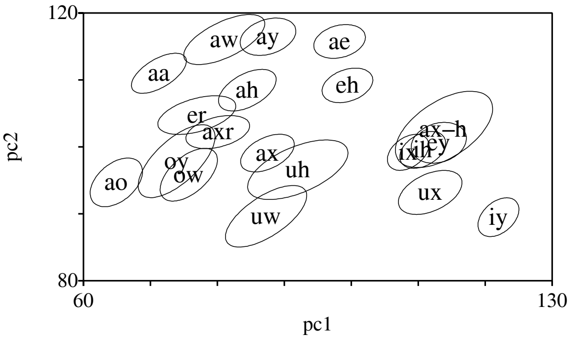

Figure 2: Average pc1-pc2 positions with 0.5-sigma ellipses of the 20 American English vowels of the 326 male speakers in the train part of the TIMIT database.

As an example, see Fig. 2 with the average positions and 0.5-sigma ellipses of the 20 American-English vowels for the 326 male speakers in the train part of the TIMIT database.

3.3.2. Discriminant analysis (DA)

Contrary to PCA, DA learns from *labeled* multi-variate data and finds subsequent dimensions for which the classification is maximal. The measure used for classification is the quotient of between-classes-variance and within-classes variance, hence the requirement for labeled data. In Linear DA (LDA) one, over all classes pooled, covariance matrix is used in the distance calculation. In Quadratic DA (QDA) class-specific covariance matrices are used in the distance calculation. In this case the boundaries between classes, when projected on a two-dimensional space, are quadratic curve segments, while they are straight line segments for LDA.

3.3.3. Procrustes normalization

The Procrustes transform (PT) allows for a transformation of one data set to match another data set as closely as possible, while the structure of the transformed data is preserved, thus leaving all relative distances between the data points intact. The only admissible operations are thus dilation, translation, rotation and reflection.

3.3.4. Feedforwar neural nets (FNN)

Contrary to DA, neural net models do not make any assumption on the distribution of the data, i.e. these do not have to be normally distributed. FNN’s have separate learning and classification phases and therefore belong to the class of *supervised* models. In the learning phase, stimuli are presented one-by-one and the network adapts to the individual stimulus. This contrasts with discriminant learning which is batch-oriented, i.e. a statistic like the mean is derived from all data taken together. The forthcoming thesis of Weenink [13] discusses several aspects of FNN, such as topology, capabilities of one-, two-, and three-layer nets, non-linearities, and several cost functions.

3.3.5. Adaptive resonance theory networks (ART)

These ART neural network models are based on the perception theories of Grossberg, an overview of this theory can be found in [3]. ART is based on *unsupervised* real-time learning. An important problem to be solved in this context is the *stability-plasticity* dilemma: how can one learn new things without gradually forgetting old things. A predictive variant of ART called CategoryART was developed by the second author. It is able to learn to predict a prescribed category given a prescribed *n*-dimensional input vector. CategoryART shows excellent performance on a benchmark neural network test set, the artificial two spirals problem [13], but much poorer performance on our real-world problem of recognizing overlapping vowel categories.

4. Adaptive vowel recognition

In this section we will present several results of applying vowel recognition and speaker adaptation procedures to various data sets. First we will present in sect. 4.1 some extrinsic speaker normalization methods and then in sect. 4.2 one intrinsic method.

4.1. Extrinsic speaker normalization

Any extrinsic method has preliminary knowledge about the data of a speaker, that then can be used to transform, adapt or normalize the data of that speaker or the model to be applied before starting vowel recognition. Already in 1967, Pols and colleagues [8,9,10] clearly demonstrated that level normalizing followed by 'centering' the formant or band-filter vowel data per speaker, clearly improved recognition performance. This *centering* implies that all centers of gravity (average position of the 12 vowels per speaker) were put in the origin. However, there exist several other normalization methods.

4.1.1. Bias adaptation in a neural net for formant data

Another way of extrinsic speaker normalization can be achieved by adapting only the bias of the hidden or the output layer per speaker of an already trained neural net.

We used here for training the log-transformed formant data of the 50 male Dutch speakers [9]. Since these formant data contain no durational information, we reduced the 12 vowel categories to 9, by combining the short-long vowel pairs /O-o:/, /Y-2:/, and /I-e:/ (in SAMPA notation). The topology of the neural net was (3, *h*, 9): 3 input nodes (log F1, log F2, log F3), *h* hidden nodes (varying from 3 to 7), and 9 output nodes (the 9 Dutch vowel categories).

Table 3: Vowel classification and bias adaptation performance of a neural net with topology (3, *h*, 9), trained with the grouped 50 male formant set MG50 and tested with various other sets. For more details, see text.

H	Test set	Test (%)	H_{all} (%)	H_{ind} (%)	O_{all} (%)	O_{ind} (%)
3	MG10	91.7	92.5	100	95.8	98.3
3	WG10	74.2	91.6	96.7	73.3	76.7
3	CG10	26.7	90.8	98.3	39.1	36.7
	Mean	64.2	91.6	98.3	69.4	70.5
3	MG50	89.3	89.3	96.7	89.3	94.7
3	WG25	78.3	88.0	97.3	83.3	86.7
7	MG10	93.3	94.1	99.2	95.0	96.7
7	WG10	71.7	82.5	85.8	77.5	81.7
7	CG10	51.7	63.3	70.8	58.3	62.5
	Mean	72.2	80.0	85.3	76.9	80.3
7	MG50	93.2	93.2	98.0	93.2	94.5
7	WG25	75.6	84.6	92.0	85.0	83.7

Table 3 shows the classification and adaptation performance in terms of percentage vowels correct for various test sets, varying from the 10 male (MG10), 10 female (WG10) and 10 children (CG10), to the 50 male (MG50) (that was also used for training), and the 25 female speakers (WG25) [10]. Test results are presented in the third column for 3 and 7 hidden nodes (without bias adaptation) and then in the subsequent columns also for bias adaptation of the *hidden* (columns 4 and 5) or the *output* nodes (columns 6 and 7). The subscript 'all' indicates that the adaptation of the biases is done for all speakers together, whereas the subscript 'ind' indicates that this is done for every speaker individually.

The general tendency for the test sets (column 3) is that classification performance gets worse when the speaker category goes from men to women and then to children, irrespective of the number of hidden nodes. Generally, classification performance increases when the number of hidden nodes increases, for instance MG50 (89.5 -> 93.1%), MG10 (91.7 -> 93.3%), or CG10 ((26.7 -> 51.7%). However for the women speakers (WG10 and WG25) there is a slight decrease. When the hidden nodes are allowed to adapt themselves to the complete test sets, we see a remarkable increase in percentage correct. With a change in only 3 bias parameters, WG10 improves from 74.2 to 91.6%, for CG10 the increase is even more extreme from 26.7% to 90.8%. By adapting the biases to each individual speaker (column H_{ind}) the results get even better. The recognition scores for this case came rather close to 100% for all speaker categories, which is quite extraordinary if we realize that the training was performed with the data from the male speakers. The last two columns (bias adaptation of the output layer) show high scores when the speaker category of the test set equals the speaker category of the training set (men). The improvement is only small when the speaker categories differ.

For these formant data, bias adaptation of the *hidden* layer works best, indicating that a simple translation of the hyperplanes is sufficient to guarantee proper adaptation. At the same time it is powerful enough even for the adaptation of vowels spoken by children to those spoken by men. For similar results while training with formant data from 25 female speakers, see [13].

The results in this section are extremely encouraging, however, they were based on idealized formant data, implying a manual segmentation and manual formant measurements. Therefore we wanted to run these tests again, but this time on automatically derived bandfilter data. Unfortunately, the audio files of the 50 male and 25 female Dutch vowel data are no longer accessible to us, so we ran the test on the TIMIT database, which also implies substantially more, and more natural, data. Before presenting these results in sect. 4.1.3, we will first present another normalization procedure on these TIMIT bandfilter data, by using discriminant analysis.

4.1.2. Discriminant analysis on TIMIT bandfilter data

The bandfiltered vowel data from the TIMIT database are a real challenge in many respects. The database is much bigger (78,374 segments, see Table 2) and much more variable (both stressed and unstressed vowels from many male and female speakers in several dialects), the bandfilter data are automatically derived and the TIMIT data are widely accepted as a standard data set [4,6,7]. We used a filterbank with 18 filters 1 Bark wide and also 1 Bark apart, as implemented in *praat* [2,13]. The bandfilter analysis is performed on three frames in each vowel segment: a 20-ms frame at the midpoint and two frames at 25 ms before and after the midpoint. Because the 20 vowels did not occur equally often per speaker, we calculated the average representation for each vowel for each speaker. These 20 average (so-called *summary*) vowels per speaker were the basis for an unbiased determination of for instance the PCA representation in Fig. 2 above and for the discriminant classifier described next. For these tests we used the summary data of 630 speakers, in principle this results in 630 x 20 = 12,600 entries. However, since sometimes a speaker did not produce a certain vowel at all, the actual number was 11,579 (see Table 2).

Table 4 shows the vowel recognition performance for these 18-dimensional bandfilter data of a discriminant classifier trained with three different summary data sets and tested with the same three sets. First the total set of male plus female data (MF-S), then the 8,078 only-male entries (M-S) and the 3,501 only-female entries (F-S). For instance, the row marked MF-S in the table reads as follows: a discriminant classifier trained on the whole data summary set (MF-S, male plus female data) shows 57.5% correct when tested on the same data set. It shows 63.4% correct when tested on the male data set only (M-S) and 44.0% correct on the female set only (F-S). The increase from 57.5% for MF-S to 63.4% for M-S and the following decrease to 44,0% for F-S is due to the fact that MF-S training is biased, since that set contains many more male than female data (see Table 2). Having separate classifiers for male (M-S) and female (F-S) data, substantially improves the percentage correct to 66.3 and 62.4, respectively. At the same time these classifiers show worse performance when the "other" set is tested, as the numbers 23.3 and 25.9 show, but see sect. 4.1.5. These differences between the speaker groups still remain when we perform a speaker normalization (see below), as the numbers between parentheses show.

Table 4: Percentage correct vowel scores with a discriminant classifier. Between parentheses the scores after speaker normalization. For more details see text.

	tested with			
trained with	MF-S	M-S	F-S	Entries
MF-S	57.5 (64.5)	63.4 (70.4)	44.0 (51.0)	11,579
M-S	53.3 (59.6)	66.3 (73.4)	23.3 (28.0)	8,078
F-S	37.0 (39.8)	25.9 (26.7)	62.4 (70.2)	3,501

In the straightforward *speaker normalization procedure* that we have applied here, we have corrected the bandfilter data by the difference of the speaker's average and the group average. We see that this improves the correct scores substantially, but that the differences between the M-S and F-S sets remain.

4.1.3. Neural net bias adaptation for bandfilter data

After having shown above the possibilities of discriminant analysis for vowel identification, applied to the TIMIT bandfilter data plus a simple form of speaker normalization, we want to see next what bias adaptation can achieve in a neural net. We were quite optimistic about that given the excellent result achieved in sect. 4.1.1 for formant data.

Input data are 18-dimensional bandfilter spectra of all vowel segments in TIMIT. However, since we preferably wanted the training time for the neural nets to be short, we reduced the dimensionality of the input data to 9 dimensions by applying a principal components analysis. These first 9 factors already comprise 97.3% of the total variance. The topology of the feedforward neural nets was (9, *h*, 20): the first 9 principal components as input, 1 to 9 hidden nodes, and the 20 American English vowel classes as output. We trained the neural nets (batch training with maximally 1,000 epochs) with either male or female summary data (see Table 2). The test data will always be independent from the train sets.

Once the neural net has learned the data set, we perform the adaptation process as follows:

- The data in the test set are grouped according to speaker identity;
- We make a copy of the trained neural net and arrange its parameters in such a way that only the biases of the hidden or the output layer are allowed to be modified during adaptation. We started the minimization in the adaptation step with the biases as they resulted from the training phase, since initialization with random numbers got too often stuck in local minima;
- We select the vowel spectra for a particular speaker and again train the neural net. However, this time *only* the selected biases are allowed to change. Since much less data are involved here (maximally 20 spectra) and since only a few bias weights have to be modified, we limited the number of epochs to 100;
- We then use the neural net as a classifier for the data of the selected speaker and we record the fraction correct;
- We then move on to the data of the next speaker until all test speakers have been processed.

This procedure was repeated for each of the 9 different neural net topologies, and the whole process was repeated ten times.

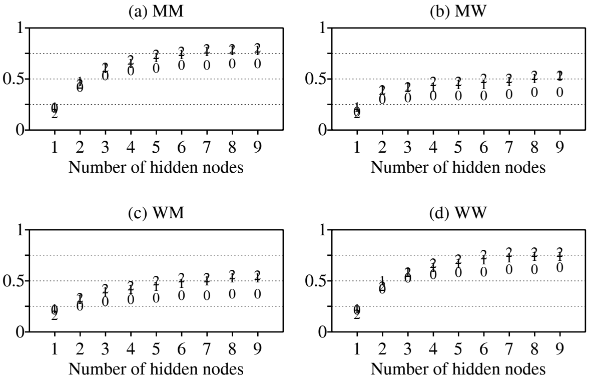

Figure 2: Vowel recognition results for TIMIT bandfilter data with bias adaptations in a neural net. For more details, see text.

Figure 2 summarizes the adaptation results. In each of the four panels the fraction correct vowel classification is displayed as a function of the number of hidden nodes. A '0' indicates the result without speaker adaptation, the symbols '1' and '2' show results for speaker adaptation of the biases in layer 1 and 2, the hidden and the output layer, respectively. With the two symbols M (men) and W (women) it is indicated what were the training data and what were the test data. Thus in parts (a) and (d) we have a learning task in which the train and the independent test data belong to the same speaker group, either MM or WW. Whereas in parts (b) and (c) the train and test data belong to different speaker groups (MW and WM). We note in all four panels that adaptation is effective and shows larger fractions correct than the baseline condition ‘0’. Contrary to the results with formant data (see sect. 4.1.1), this time the bias adaptation of the *output* layer ‘2’ appears to be slightly more effective than the adaptation of the hidden layer ‘1’. For all three conditions ‘0’, ‘1’ and ‘2’ the fractions correct gradually increase and then level off. The asymptotic scores for the best condition ‘2’ in the four panels are: MM 80.0%, MW 52.9%, WM 55.0% and WW 77.8% correct.

Just as for formant data (see sect. 4.1.1), the bias adaptation model also seems to work very well for bandfilter spectra, especially when the adaptation is to a speaker from the same male or female group. There was also adaptation when train and test sets belonged to different speaker groups, but in these cases the absolute fractions correct were not impressive. Apparently the differences between male and female vowel spectra cannot be annihilated by simply adapting the biases, but see sect. 4.1.5. In the next section we will test a, potentially more powerful neural net model to help us to find a better speaker adaptation model.

4.1.4. Vowel recognition with CategoryART

We trained this CategoryART net with the male train set of 6,008 summary vowel entries, and we used the same 9-dimensional input data as before. The independent test set consist of all 13,889 male entries, see Table 2 for more details. The *stability-plasticity* dilemma (see sect. 3.3.5) was tackled by varying ρ and β. The vigilance parameter ρ was varied in 7 steps from 0.7 to 1 and the learning rate parameter β was varied in 5 steps form 0.05 to 1. These 35 different parameter combinations were tested with match track *on* as well as *off*. When *on* the network artificially increases the vigilance level until a matching node has been found (or created). This actually implies that 2 x 35 different CategoryART’s have been trained. The tests then were repeated 10 times and the results averaged. The results for match track *off* are displayed in Figure 3.

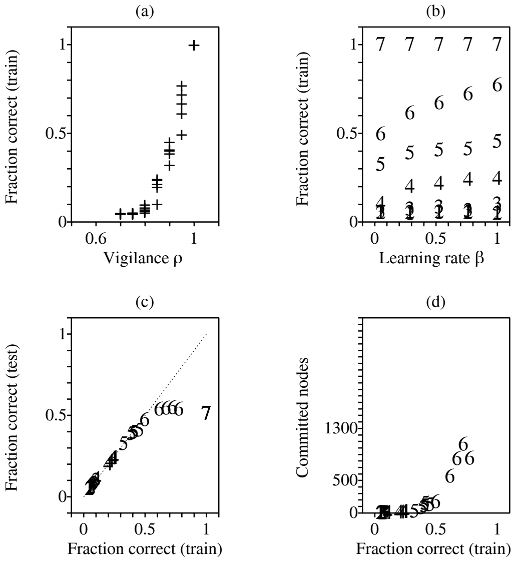

Figure 3: Summary of the CategoryART training and testing results with TIMIT vowel data. For more details, see text.

In panel (a) the fraction correct is displayed as a function of the vigilance parameter for the train set. We note a steady increase, leading to a correct classification of all items in the train set when ρ=1. Panel (b) shows the fraction correct as a function of the learning rate for various vigilance levels labeled 1 (ρ=0.7) to 7 (ρ=1) for the train set. Apparently the vigilance parameter has a much larger effect on the fraction correct than the learning rate parameter. Panel (c) shows the generalizing properties of the network. The scatter plot shows the fraction correct for the test set *versus* the train set at the seven different vigilance levels. The plot clearly shows that for the first five vigilance levels the performance on the train set and the test set are almost equal and increase monotonously with vigilance level. At vigilance level 6 where ρ=0.95, the fraction correct levels off to a value of roughly 0.52. The value of 0.52 seems to be the maximum performance on the test set. Panel (d) shows the number of committed nodes as a function of the fraction correct for different values of the vigilance parameter. We have left out the 7th level since the number of committed nodes for ρ=1 equals the number of items in the train set: 6,008.

When we compare these fractions correct with the numbers displayed in Table 4 for the discriminant classifier, we must conclude that the generalization capabilities of the exemplar-based CategoryART network are not too impressive with this kind of (overlapping) data. For example, the discriminant classifier scores 66.3% correct on the (summary) test set which is clearly better than our 52%. Also the feedfoward neural network shows better results as was shown by the ‘0’ symbols in Fig. 2 when the number of hidden nodes exceeds 4 or 5.

4.1.5. Procrustes normalization

In Table 4 we showed identification results for a discriminant classifier that were quite good within one type of speakers (either male or female), but much less so for differing speaker groups (M-S *vs*. F-S and F-S *vs*. M-S). By using the Procrustes transform to optimally transform the average 18-dimensional female data to the male data and then applying that transform to all individual female data, we can improve the 23.3% score (for the condition: trained with M-S and tested with F-S) to 58.3% and similarly for transforming the average 9-dimensional female male data to the male data, the score improves from 25.9% score to 57.3%.

4.2. Intrinsic speaker normalization; static/dynamic data

Various extrinsic speaker normalization procedures have been presented above. However, always information about the whole vowel set of one or more speakers was required for that. A truly successful and more human-like adaptation method should have to adapt faster without requiring at least one item for each vowel class. One solution could be *to make local differences have global consequences*. We have implemented such a model in the following way:

- we start with a trained discriminant analyzer and suppose that an average reference vowel position is available for each vowel;
- then an unknown vowel is presented to the classifier and the distance to each reference vowel is determined.

Table 5: Static (one central frame) and dynamic (3 frames) discriminant classification of the TIMIT vowels with 18-dimensional bandfilter data. Percentages correct are presented for various data sets and various train and test conditions. The numbers between parentheses concern scores after a simple form of speaker normalization. For more details, see text.

			tested with			
trained with			Static		Dynamic	
Data	# segments	type	Train data	Test data	Train data	Test data
All	40,468	M	47.3 (50.3)	48.4 (50.6)	56.0 (58.4)	55.7 (58.1)
	16,995	F	46.4 (49.3)	46.1 (48.6)	56.3 (59.0)	54.7 (57.1)
All Summary	6,008	M-S	65.9 (76.6)	66.8 (77.3)	84.4 (89.9)	83.6 (89.0)
	2,490	F-S	62.0 (73.9)	61.2 (73.6)	83.8 (88.7)	79.7 (86.6)
All Stressed (+)	25,706	M +	49.8 (53.0)	50.9 (53.7)	60.7 (63.1)	60.0 (62.7)
	10,622	F +	49.0 (51.9)	48.9 (51.3)	60.7 (63.5)	59.4 (62.3)

- the shortest distance is determined and tested against a tolerance criterion, because we want to make sure that the unknown is “close enough” to the reference.
- If the “close enough” criterion is satisfied, *all* references are moved parallel to the direction of the difference vector, otherwise no changes are made.

Before applying this model, we present first in Table 5 the basic results of discriminant classification for various data sets. On the left-hand side the training data are specified, whereas on the right-hand side the results for various test sets are given. We distinguish (see also Table 2) ‘all data’ (train set 57,463 segments; test set 20,911), ‘all summary data’ (train 8,498; test 3,081) and ‘all stressed vowel data’ (train 36,328; test 13,234), each time for male and female speakers separately. Furthermore, we distinguish between ‘static’ and ‘dynamic’, indicating that either one central frame or three frames per vowel segment are used (see also sect. 4.1.2). Finally, the numbers between parentheses concern correct scores after speaker normalization. This is the same straightforward procedure as applied before in sect. 4.1.2, namely correcting the individual bandfilter data by the difference of the speaker’s average and the (male or female) group average.

For instance the last row in this table should thus be read in the following way: A discriminant classifier trained with the 10,622 stressed vowels of 136 female speakers in the *train* part of TIMIT shows a correct score of 49.0% when tested with the same data set, and shows a correct score of 48.9% when tested with the 4,389 stressed vowels of 56 female speakers in the *test* part. When three frames, rather than one frame of bandfilter data are used for training and testing, these scores are 60.7 and 59.4%, respectively. The higher scores between parentheses concern the above mentioned simple form of speaker normalization. It is furthermore clear that the stressed vowels alone have a consistently higher score than all vowels, even though ‘word stress’ is based on the normative transcription in the TIMIT pronunciation lexicon, rather than on actually realized word stress.

If we now apply the earlier described more human-like adaptive procedure for speaker normalization, we achieve the results as presented for the male summary data in Table 6.

The parameter α defines how much the current reference positions will move. With α=0 there is no change, so this condition then is identical to the male summary data condition (66.8%) in Table 5. Whereas for α=1 the difference between the reference and the unknown input is completely corrected for, however, this can easily lead to an unstable system that jumps from one input to the next input. The distinction between *blocked* and *mixed* has to do with the order in which the data are presented to the classifier: either all vowel data from one speaker sequentially (blocked) or each new vowel item from a different speaker (mixed). Clearly these are very challenging conditions. Speaker normalization can only rely one (unlabeled) item at a time. We see that in the blocked condition some adaptation is effective, with a maximum score of 69.7% correct for α=0.2. However, in the mixed condition no real adaptation appears to be possible. Still, there is always a difference between the scores under both conditions, thus reflecting the same trend as in human vowel recognition (as described in sect. 2).

Table 6: Vowel classification scores for the 18-dimensional *male* bandfilter summary data from TIMIT for a *blocked* or *mixed* presentation of the data. *Test* data (112 males) are independent of the *train* data (326 speakers). Each cell in the column *mixed* is the average of 10 different randomized test sets.

α	Blocked	Mixed	Difference
0.0	66.8	66.8	0.0
0.1	69.0	66.1	2.8
0.15	69.0	64.8	4.2
0.2	69.7	64.9	4.8
0.3	68.9	64.5	4.3
0.4	67.8	62.8	5.0
1.0	59.0	52.7	6.3

5. Discussion and conclusions

By systematically analyzing various vowel data sets (Dutch 10 male, 10 female, 10 children; Dutch 50 male and 25 female; American English male and female vowel segments from TIMIT), with several different parameterizations (formant data and 18- and 9-dimensional bandfilter data), with several pattern recognition procedures (discriminant analysis, feedforward neural nets, and adaptive resonance theory networks) and several extrinsic speaker normalization procedures (centering, neural net bias adaptation, CategoryART, and Procrustes normalization) and one intrinsic one (gradual change in references), we have learned a lot about the possibilities of acoustic vowel recognition. We were even able to imitate some human vowel recognition behavior, in terms of the distinction between scores for a blocked and a mixed condition. It also became abundantly clear that a straightforward comparison of percentage correct scores is not so simple, because many dependencies exist.

For instance, if hand-corrected formant data are available, the correct scores can be very high and bias adaptation in a neural net can be very efficient. However, if we have to work with more realistic, and automatically derivable, bandfilter data, all scores are substantially lower. A discriminant classifier (that however requires labeled data) generally performs very well, see the older data of Pols and colleagues [8,9,10], as well as Table 4 for our 9-dimensional bandfilter data and Table 5 for our 18-dimensional bandfilter data.

Neural net bias adaptation is another interesting approach, because it is appealing as a human analagon: once the network is trained, only small adaptations of a small set of bias parameters suffices to adapt to a new speaker. However, the substantial differences between male and female bandfilter data cannot be overcome with this bias adaptation alone. For that a Procrustes transform appears to be a better solution (see sect. 4.1.5).

Conceptually the idea behind ART is attractive but its performance appears to be rather poor for our vowel data in which the categories are partly overlapping (see sect. 4.1.4).

The intrinsic speaker normalization procedure that we developed, and that is perceptually relevant and is based on the concept of local differences having global consequences, does perform according to expectations (see Table 6) but the benefits compared to discriminant analysis perse, are small and even negative in the mixed condition.

Our analyses also showed that frequently it was advantageous to train with the so-called summary data, implying that some of the within-speaker variability as well as the unbalanced vowel distribution of frequency of occurrence, was taken out of the data by using average positions per vowel.

It was also shown in sect. 4.2 that, using a simplified form of dynamic information (three frames rather than one central frame per vowel segment), significantly improved vowel recognition performance. Using even more detailed spectro-temporal information will probably improve further the performance, however, not all pattern recognition procedures can easily handle temporal data.

It should not worry the reader that the percentage correct scores for the TIMIT data that we present, are generally lower than found elsewhere in the literature. This is fully based on the fact that we limited ourselves mainly to static acoustic information only, thus neglecting context and language modeling which will certainly improve overall performance.

6. References

[1] Adank, P., van Hout, R. & Smits, R. (2004), “An acoustic description of the vowels of Northern and Southern Standard Dutch”, *J. Acoust. Soc. Amer.*, 116: 1729-1738.

[2] Boersma, P.P.G. & Weenink, D.J.M. (1996), *Praat, a system for doing phonetics by computer, version 3.4*, Report 132, Institute of Phonetic Sciences University of Amsterdam. For updates, see http://www.fon.hum.uva.nl/praat/.

[3] Grossberg, S. (1998), *The link between brain learning, attention, and consciousness*, Techn. Rep. CAS/CNS-TR-97-018, Boston University.

[4] Halberstadt, A.K. & Glass, J.R. (1997), “Heterogeneous acoustic measurements for phonetic classification”, *Proc. Eurospeech 1997*, Rhodes: 401-404.

[5] Lamel, L., Kassel, R. & Seneff, S. (1986), “Speech database development: Design and analysis of the acoustic-phonetic corpus”, *Proc. DARPA Speech Recognition Workshop*: 100-109.

[6] Lee, K.F. & Hon, H.W. (1989), “Speaker-independent phone recognition using hidden markov models”, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37: 1641-1648.

[7] Meng, H.M. & Zue, V.W. (1991), “Signal representation comparison for phonetic classification”, *Proc. IEEE-ICASSP 1991*, Toronto: 285-288.

[8] Plomp, R., Pols, L.C.W. & van der Geer, J.P. (1967), “Dimensional analysis of vowel spectra”, *J. Acoust. Soc. Amer.*, 41: 707-712.

[9] Pols, L.C.W., Tromp. H.R.C. & Plomp, R. (1973), “Frequency analysis of Dutch vowels from 50 male speakers”, *J. Acoust. Soc. Amer.*, 53:1093-1101.

[10] Van Nierop, D.J.P.J., Pols, L.C.W. & Plomp, R. (1973), “Frequency analysis of Dutch vowels from 25 female speakers”, *Acustica*, 29:110-118.

[11] Weenink, D.J.M. (1985), “Formant analysis of Dutch vowels from 10 children”, *Proc. Inst. of Phonetic Sciences University of Amsterdam*, 9: 45-52.

[12] Weenink, D.J.M. (1986), “The identification of vowel stimuli from men, women, and children”, *Proc. Inst. of Phonetic Sciences University of Amsterdam*, 10: 41-54.

[13] Weenink, D.J.M. (forthcoming), *Speaker-adaptive vowel identification*, Ph.D thesis University of Amsterdam.

[14] Weenink, D.J.M. & Pols, L.C.W. (1999), “Multi-speaker vowel classification with adaptive neural networks”, *Proc. Int. Congress of Phonetic Sciences*, San Fransisco, Vol. 3: 1633-1636.