Adaptive Vowel Normalization: Simulating Human Recognition in Mixed and Blocked Speaker Context

David J. M. Weenink

Institute of Phonetic Sciences
University of Amsterdam, The Netherlands
David.Weenink@hum.uva.nl

ABSTRACT

We introduce a model for speaker adaptation that assumes no knowledge about speaker identity. The model is found to reproduce the difference in human vowel recognition performance for stimuli presented in blocked and mixed speaker context.

1 INTRODUCTION

Several experiments have shown that subjects, when confronted with vowel-like stimuli from different speakers, show better recognition performance when successive stimuli come from the same speaker than when the speaker identity varies unpredictably (e.g., [1], [2], [3], [4]). In the literature these conditions are called blocked and mixed, respectively. Most of the time, the mixed/blocked effect is not large, only a few percent, but the effect is consistent and statistically significant.

We have built a model that qualitatively reproduces this effect.[1] The precondition for the model is a system where (1) the centroid for each vowel is known and (2) the overall covariance matrix of the vowel space is approximately known. For the classification procedure these are the only two sources of information needed. They can easily be determined in a training session, and suffice to reproduce the mixed/blocked effect. No speaker dependent information is needed.

2 THE TEST DATABASE

The TIMIT acoustic phonetic speech corpus is a good data base for testing the speaker adaptive vowel normalization procedure because it contains labeled and segmented speech from a great number of speakers [6]. All the sound and label files in the corpus were made accessible in the *praat* program [5]. From the 22 different vowels and diphthongs that are present in the TIMIT phoneme database we have selected the 13 monophthong vowels that were also selected by [7]. These vowels are `iy`, `ih`, `eh`, `ey`, `ae`, `aa`, `ah`, `ao`, `ow`, `uh`, `uw`, `ux` and `er`. We only used the stressed vowels. Stress was determined from lexical stress by time alignment of the realized phonemes in the words that constitute a sentence and of the phonemes in the ideal pronunciation of this sentence according to the dictionary by means of a standard dynamic programming algorithm [8]. All the vowels pronounced by the 438 male speakers in both the *train* and the *test* part of TIMIT were brought together in one collection. This resulted in 35,385 vowels. We performed the following analysis steps:

- The sentences in which one or more selected vowels occurred, were marked in the database.
- An automatic band filter analysis was performed on all the marked sentences with the *praat* program. The band filtering was performed in software with a filter bank of 18 filters equally spaced on a bark frequency scale, i.e., via band filtering in the frequency domain.[2] The first filter had its centre frequency at 1 Bark and filters were spaced 1 Bark apart. The output of each filter is a value in dB's. The exact specification of the bark filters can be found in [9]. For the analysis, a window length of 25 ms and a time step of 1 ms were chosen.
- For each selected vowel, three analysis frames were chosen: one at the centre of the vowel and the two others at 25 ms before and 25 ms after the centre position. Vowel identity and speaker identity were both stored together with the analysis results for later processing. In general there were multiple replications of the same vowel by the same speaker.
- To neutralize intensity variations between vowels, the 18 band filter values in each frame were rescaled to a fixed intensity (of 80 dB).

[1] The model has been implemented by making a very small change in the discriminant classifier from the *praat* program[5].

[2] For more details see praat manual: `Sound to BarkFilter...`

ISBN 1-876346-48-5

- The vowel band filter data were collected in a `TableOfReal`-object with 35,385 rows and 54 (= 3×18) columns, where the row labels indicate vowel class.
- A Discriminant analysis was done on the `TableOfReal`-object.

To give an indication of the distribution of the vowels, we have plotted in fig. 1 the distributions with their 1σ-ellipses in the discriminant plane. This is the plane where discrimination is optimal. One clearly notices the enormous spread within each vowel class. The use of the same discriminant as a classifier[3], resulted in 59.3% correct classifications for the 13 vowel classes.

In table 1 we present the confusion matrix for this classification. In the last column, the table also gives information about the frequency of occurrence of the vowels.

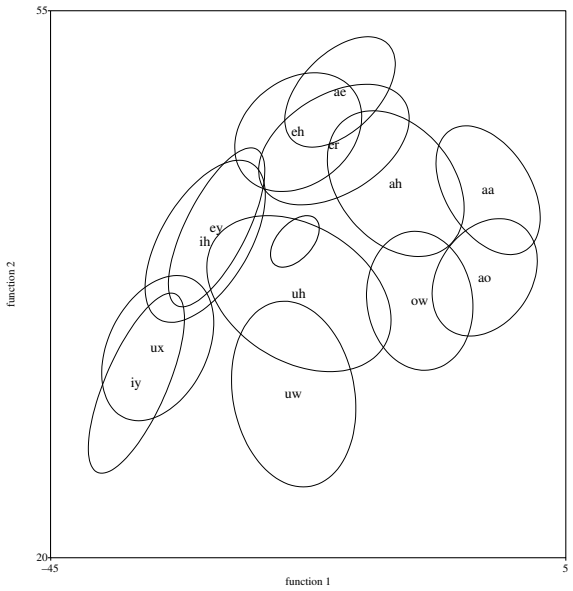

Figure 1: The distribution of the 35,385 vowels in the discriminant plane. The ellipses are the 1σ ellipses that include approximately 39.5% of the data. The vowels are from the 438 male speakers that are present in both the train and the test part of the TIMIT corpus. All eight dialect regions are represented and all vowels selected had word stress. The 1σ distribution of the 438 average spectra of the speakers is shown by the small ellipse at the centre.

3 THE ADAPTIVE SPEAKER NORMALIZATION PROCEDURE

The basis of the model is the ability to learn the joint vowel centroids from the current input. This learning proceeds as follows. A given input vector is compared with all 13 reference vectors (the vowel centroids) and the best match is chosen. When the classifier signals that the probability of group membership[4] in the match is larger than 0.5, the distance d between the input vector x and the best match reference c_k is calculated. As a result the positions of *all* 13 reference vectors are moved in the direction of the vector d by a fraction α. The new references c_i' in terms of the old references c_i will then become:

$$c_i' = c_i + \alpha d, \qquad \text{where} \qquad 1 \leq i \leq 13.$$

The next input will then be classified with respect to the modified reference system. When α equal 0 no adaptation will happen, when α equals 1 we adapt completely and with α greater than 1 we overshoot. In table 2 we show the classification results for various values of α and a minimum probability 0.5 for the data.

The scores in the cells in the mixed condition have been averaged over a number of trials. In each trial we supplied a different randomized sequence of inputs to the classifier. The table shows that for small α (e.g., $\alpha = 0.1$), the classification results in the blocked speaker condition are actually somewhat better than the results without adaptation: 60.3 % versus 59.3 %, respectively. The algorithm has actually *learned to normalize for speaker differences without knowing anything about speakers.* The table further shows that classification in the blocked condition was always superior

[3]The characteristics of the classification procedure are as follows. We perform recognition on the 18 dimensional band filter vectors with the covariance matrices of the 13 vowel classes *pooled*. When we classify with all the 13 distinct covariance matrices instead of the pooled matrix, we only get a 0.3% better classification result. Given the much larger number of parameters in the latter classifier, we prefer pooling. The pooled model uses 405 parameters: 13×18 for the means plus $18 \times (18+1)/2$ for the pooled covariance matrix. The classifier without pooling uses another 2268 parameters extra that originate from the 12 extra covariance matrices that are needed.

We also use the *a priori* probabilities. Not using a priori probabilities results in a 1.8% decrease in performance.

[4]The posterior probabilities of group membership p_j for a vector x are defined as

$$p_j = p(j|x) = \frac{\exp(-d_j^2(x)/2)}{\sum_{k=1}^{numberOfGroups} \exp(-d_k^2(x)/2)},$$

where $d_i^2(x)$ is the generalized squared distance function

$$d_i^2(x) = (x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) + \ln|\Sigma_i^{-1}|/2 - \ln(aprioriProbability_i)$$

that depends on the individual covariance matrix Σ_i and the mean μ_i for group i. When the covariance matrices are *pooled*, the squared distance function reduces to

$$d_i^2(x) = (x-\mu_i)'\Sigma^{-1}(x-\mu_i) - \ln(aprioriProbability_i),$$

and Σ is now the pooled covariance matrix. The a priori probabilities will have values that normally are related to the frequency of occurrence in the groups during the training process of the discriminant classifier.

Table 1: Confusion matrix with marginals for the 13 vowel classes. The last column in the table shows the frequency of occurrence of each vowel class and equals the sum of the elements in that row. The elements in the last row sum the responses in the corresponding column. The bottom-right element shows the total number of entries in the table and equals the sum of the elements in the last row as well as the sum of the elements in the last column. Dividing the sum of the elements on the diagonal by this number and scaling to percentages, gives 59.3% correct classification. For the classification process, covariance matrices were pooled and the a priori probabilities were used. These a priori probabilities can be derived from the last column in this table.

	aa	ae	ah	ao	eh	er	ey	ih	iy	ow	uh	uw	ux	Sum
aa	1861	113	308	399	40	66	·	3	·	71	1	·	·	2862
ae	76	2781	61	·	634	1	141	50	9	·	·	·	·	3753
ah	311	127	955	96	312	12	2	53	·	235	44	6	1	2154
ao	536	9	62	1969	5	51	2	1	1	300	3	5	·	2944
eh	52	640	335	5	1690	125	306	484	12	33	12	3	3	3700
er	10	9	27	5	110	1564	5	105	13	9	8	5	24	1894
ey	·	92	12	·	336	8	853	583	264	1	1	·	3	2153
ih	1	84	111	·	447	80	523	2145	733	40	147	21	170	4502
iy	·	11	2	·	60	21	378	855	5045	1	4	7	222	6606
ow	72	3	331	540	34	14	·	12	·	958	57	31	1	2053
uh	1	1	44	24	14	16	·	115	5	102	105	45	28	500
uw	·	1	15	14	2	17	·	27	5	75	38	279	53	526
ux	·	·	8	·	13	25	9	271	492	5	33	121	761	1738
	2920	3871	2271	3052	3697	2000	2219	4704	6579	1830	453	523	1266	35385

Table 2: Classification results with the adaptive procedure for the 35,385 vowels. Each cell in the column labeled mixed is the average of 10 trials.

α	blocked	mixed	Difference
0.0	59.3	59.3	0.0
0.1	60.3	56.7	3.7
0.2	60.1	55.2	4.9
0.5	58.6	48.1	10.5
1.0	54.4	30.6	23.8

to classification in the mixed condition. The difference between the two conditions increases when α increases: a large shift in the references probably produces incorrect results when the next input is not from the same speaker. Shifts tend to be more correlated when inputs come from the same speaker.

4 CONCLUSION

We have shown that a rather simple model that adapts to an incoming stimulus, actually learns to normalize for speaker differences without having any specific information about individual speakers or even about a change in speaker context. The only precondition is that stimuli from speakers are presented in a blocked condition. As a side effect, the model shows a difference in recognition performance between stimuli in blocked and mixed speaker context.

In future experiments we will test whether these conclusions will hold when we introduce other test environments. We are thinking about the separation of train and test sets. In a variant of these tests we will use a train set with vowels produced by male speakers and a test set with vowels produced by female speakers and vice versa. Another possibility would be to have one extra adaptation in the algorithm: instead of moving all references at the same time along the same difference vector by the same amount α, we could try to adapt the reference for the vowel that matches best somewhat faster than the other references. This would result in an adaptation at possibly two different speeds.

ACKNOWLEDGEMENT

The author wishes to thank Louis Pols for his constructive comments during this study.

REFERENCES

[1] W Strange, R R. Verbrugge, D P. Shankweiler, and T R. Edman, "Consonant environment specifies vowel identity," *J. Acoust. Soc. Am.*, vol. 60, no. 1, pp. 213–224, 1976.

[2] M. J. Macchi, "Identification of vowels spoken in

ISBN 1-876346-48-5

isolation versus vowels spoken in consonantal context," *J. Acoust. Soc. Am.*, vol. 68, pp. 1636–1642, 1980.

[3] P F. Assmann, T M. Nearey, and J T. Hogan, "Vowel identification: Orthographic, perceptual, and acoustic aspects," *J. Acoust. Soc. Am.*, vol. 71, pp. 975–989, 1982.

[4] D J. M. Weenink, "The identification of vowel stimuli from men, women, and children," *Proceedings of the Institute of Phonetic Sciences University of Amsterdam*, vol. 10, pp. 41–54, 1986.

[5] P P. G. Boersma and D J. M. Weenink, "Praat, a system for doing phonetics by computer, version 3.4," Report 132, Institute Of Phonetic Sciences University of Amsterdam (up-to-date version of the manual at `http://www.fon.hum.uva.nl/praat/`), 1996.

[6] L.F. Lamel, R.H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus, saic-86/1546," in *Proc. DARPASpeech Recognition Workshop*, 1986, pp. 100–109.

[7] H. M. Meng and V. W. Zue, "Signal representation comparison for phonetic classification," in *IEEE Proc. ICASSP, Toronto*, 1991, pp. 285–288.

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

[9] A. Sekey and B.A. Hanson, "Improved 1-Bark bandwidth auditory filter," *J. Acoust. Soc. Am.*, vol. 75, pp. 1902–1904, 1984.

ISBN 1-876346-48-5