Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 24 (2001), 117–123.

VOWEL NORMALIZATIONS WITH THE TIMIT ACOUSTIC PHONETIC SPEECH CORPUS

David Weenink

Abstract

In this paper we present preliminary results of speaker normalization procedures that were tested with all 35,385 stressed vowels of 438 male speakers in the TIMIT speech corpus. First we investigate a procedure to reduce the variance in vowel space. This procedure knows about the identity of the speaker. In the next part 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

The TIMIT acoustic phonetic speech corpus is a good data base for testing vowel normalization procedures because it contains labeled and segmented speech from a great number of speakers (Lamel et al., 1986). All sound and label files in the corpus were made more accessible by us in the *praat* program (Boersma & Weenink, 1996). In a previous paper (Weenink, 1996) we reported about adaptive vowel normalization with a feed forward neural net. In this paper we will use classical linear discriminant analysis as a classifier.[1] In the current investigation we were interested in exploring to what extend vowel classification could be improved by incorporating knowledge about the speaker in the classification process.

2 Vowel selection procedure

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 Meng & Zue (1991). These vowels are iy, ih, eh, ey, ae, aa, ah, ao, ow, uh, uw, ux, er. We 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 the phonemes in the ideal pronunciation of this sentence according to the dictionary by means of a standard dynamic programming algorithm (Weenink, 1996). 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 steps:

[1] Linear discriminant analysis has been implemented in the *praat* program, see (Weenink, 1999).

- 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 Sekey & Hanson (1984). 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).

- The vowel band filter data were collected in a `TableOfReal`-object with 35,385 rows and 54 ($= 3 \times 18$) columns.

3 Variance reduction

To get an indication of the distribution of the vowels in the static raw condition (see below), 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. Using 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.

In order to reduce the spread in the data we have treated the data in the following ways:

raw The raw material, normalized only for intensity variations, consists of 18-dimensional band filter spectra B_{ijk}, where the index i ($1 \leq i \leq 13$) represents the vowel type, the index j represents the speaker ($1 \leq j \leq 438$) and k represents one of the replications of this vowel by the same speaker (k varies between 1 and 25). As one would have guessed from table 1, the maximum number of replications occurs for the vowel iy. The average number of replications is 6.2 ($= 35,385/(438 \times 13)$).

[2] See praat manual: Sound to BarkFilter...

[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.

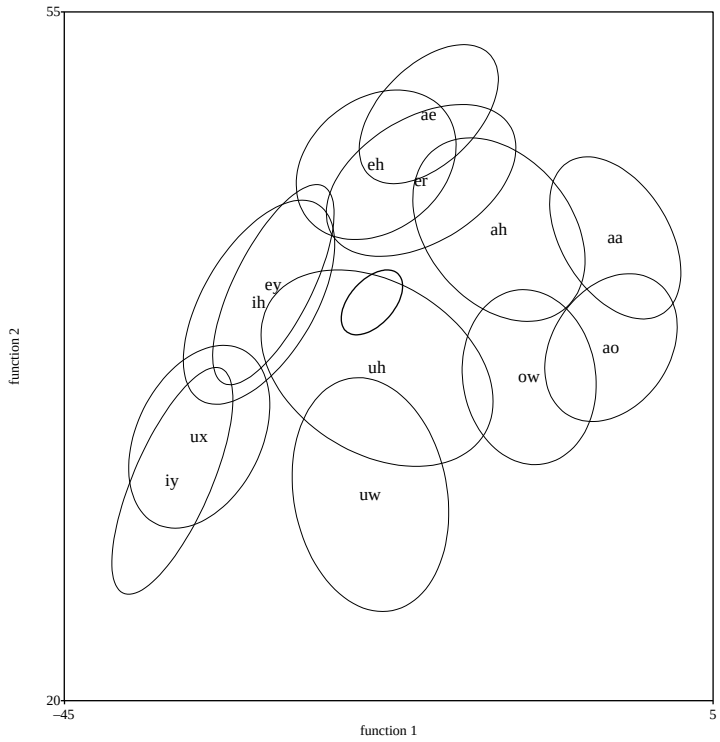

Fig. 1. The distribution of the 35,385 vowels in the discriminant plane. The ellipses are the 1σ ellipses that include approximately 39.5% of the data. The vowels are from the 438 male speakers that are present in both the train and the test part of the TIMIT corpus. All eight dialect regions are represented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

cograw The raw material, corrected for the between-speaker variance. From the raw material B_{ijk}, we calculate the normalized spectra B'_{ijk} as:

$$B'_{ijk} = B_{ijk} - (B_{.j.} - B_{...}),$$

where $B_{.j.}$ is the average spectrum for speaker j and the averaging is performed over all the speaker's different vowels and their replications, and where $B_{...}$ is the spectrum averaged over all speakers, vowels and replications. The net effect is a kind of *centre of gravity* correction.

ave Instead of multiple replications of a vowel by each speaker, we reduce the data to one exemplar per vowel by averaging over all replications of that vowel for that speaker. This operation reduces the number of spectra with almost a factor of 7 to 5374. This does not equal 438×13 because not all speakers produced all 13 different vowels at least once. Keep in mind that per speaker only 10 sentences were available).

The average spectra B''_{ij} are calculated as:

$$B''_{ij} = B_{ij.},$$

where $B_{ij.}$ is the spectrum for vowel type i from speaker j averaged over all replications.

Table 1. Confusion matrix with marginals for the 13 vowel classes obtained from the raw data. 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 discriminant functions. The first column, labeled *Condition* represents the treatment of the data as is explained in the text. The second column contains the number of band filter spectra used in the classification. The columns labeled *Static* and *Dynamic* show percentages correct classification. In the former column only the centre frame was used for the classification, in the latter column all three analysis frames were used.

Condition	# Items	Static	Dynamic
raw (B_{ijk})	35385	59.3	66.9
cograw (B'_{ijk})	35385	62.2	69.2
ave (B''_{ij})	5374	78.9	90.1
cogave (B'''_{ij})	5374	87.9	94.5

cogave The ave data corrected for the between-speaker variance. The spectra B'''_{ij} are calculated as:

$$B'''_{ij} = B''_{ij} - (B''_{\cdot j} - B''_{\cdot\cdot}).$$

Besides the normalizations as discussed above, we also introduced another source of information: *static* versus *dynamic* spectra. For the static spectrum we used the spectrum measured at the centre of the vowel (a vector with 18 numbers). For the dynamic spectra we used all three band filter spectra (at 25 ms before the centre, at the centre and at 25 ms after the centre: a vector with 54 numbers). We have calculated separate discriminant functions for the data under these eight conditions and in table 2 we present the classification results. Again the individual covariance matrices were pooled.

From this table we clearly see several trends:

- Including dynamics improves the classification process. The classification results for the dynamic spectra are always better than those for the corresponding static spectra.

- Applying speaker normalization by reducing between-speaker variance always results in better classification. This can be seen for the raw data by comparing the row labeled raw versus the row labeled cograw and for the speaker-averaged data by comparing the rows labeled ave and cogave. The effect is greater for the speaker-averaged data.

- Reducing the within-speaker variance has the greatest impact on classification. We see a dramatic increase in percentage correct when we compare the conditions raw and ave. This is in line with ANOVA results for TIMIT from Sun & Deng (1995), who find that the variance component due to within vowel variation because of different phonetic contexts is much larger than the variance due to variation among speakers. In their study they conclude that of the total variation approximately 34% is explained by differences between the phoneme units, 28% by variations within the phoneme units and 12% by variations among the speakers.

Our data show that, given the right amount of context information, classification can be significantly approved.

4 An adaptive speaker normalization procedure

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 very often (e.g. Strange et al. (1976), Macchi (1980), Assmann et al. (1982), Weenink (1986)). In the literature the conditions above are often 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.[4] 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, they are enough to reproduce the mixed/blocked effect. No speaker dependent information will be used.

The basis of the model is that it tries 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[5] in the match is larger

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

[5]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

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

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

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

Table 3. Classification results with the adaptive procedure described in section 4 for the 35,385 vowels in the raw condition. 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

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 3 we show the classification results for various values of α and a minimum probability 0.5 for the raw 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 $\alpha = 0.1$, the results for the blocked speaker condition is actually better than for the comparable raw condition in table 2: 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 to classification in the mixed condition. The difference between the two conditions increases when α increases: making a large shift in the references may be incorrect when the next input is not from the same speaker. Shifts tend to be more correlated when inputs come from the same speaker.

5 Conclusion

We have shown that when we reduce intra-speaker variance very good recognition rates for vowels can be obtained. Adding dynamic information about the vowel by just adding two measurement points left and right of the central value, further enhances recognition. We have shown also that a rather simple model that adapts to an incoming stimulus has actually learned to normalize for speaker differences without having any specific information about individual speakers or even about a change in speaker context. The only precondition was that stimuli from speakers are presented in a blocked condition. As a side effect, the model automatically 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

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.

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.

Acknowledgment

The author wants to thank Louis Pols for his critical review and constructive comments during this study.

References

Assmann, P. F., T. M. Nearey & J. T. Hogan (1982): "Vowel identification: Orthographic, perceptual, and acoustic aspects", *J. Acoust. Soc. Am.* **71**: 975–989.

Boersma, P. P. G. & D. J. M. Weenink (1996): *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/`).

Lamel, L., R. Kassel & S. Seneff (1986): "Speech database development: Design and analysis of the acoustic-phonetic corpus, saic-86/1546", in *Proc. DARPA Speech Recognition Workshop*, 100–109.

Macchi, M. J. (1980): "Identification of vowels spoken in isolation versus vowels spoken in consonantal context", *J. Acoust. Soc. Am.* **68**: 1636–1642.

Meng, H. M. & V. W. Zue (1991): "Signal representation comparison for phonetic classification", in *IEEE Proc. ICASSP, Toronto*, 285–288.

Sekey, A. & B. A. Hanson (1984): "Improved 1-Bark bandwidth auditory filter", *J. Acoust. Soc. Am.* **75**: 1902–1904.

Strange, W., R. R. Verbrugge, D. P. Shankweiler & T. R. Edman (1976): "Consonant environment specifies vowel identity", *J. Acoust. Soc. Am.* **60**: 213–224.

Sun, D. X. & L. Deng (1995): "Analysis of acoustic-phonetic variations in fluent speech using TIMIT", in *IEEE Proc. ICASSP, Detroit*, 201–204.

Weenink, D. J. M. (1986): "The identification of vowel stimuli from men, women, and children", *Proceedings of the Institute of Phonetic Sciences University of Amsterdam* **10**: 41–54.

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

Weenink, D. J. M. (1999): "Accurate algorithms for performing principal component analysis and discriminant analysis", *Proceedings of the Institute of Phonetic Sciences University of Amsterdam* **23**: 77–89.