Reduction of speaker-related variability in Polish vowel spectra

Wiktor Jassem
Institute of Fundamental Technological Research,
Polish Academy of Sciences
wjassem@amu.edu.pl

Mirosław Krzyśko
Faculty of Mathematics and Computer Science,
Adam Mickiewicz University
& Institute of Mathematics,
University of Zielona Góra
mkrzysko@amu.edu.pl

Waldemar Wołyński
Faculty of Mathematics and Computer Science,
Adam Mickiewicz University
wolynski@amu.edu.pl

Sources of variability in the speech signal and their statistical treatment are discussed. In the experiment, a total of 3390 vowel spectra obtained from a specially constructed 100-word list read by five male and five female speakers were measured, described in terms of five variables, F_1 F_2, F_3, F_4, F_5 and subjected to statistical analysis. First, the 5D data are normalized, and then treated using a novel discriminant analysis based on a multi-stage comparison of pairs of classes (vowel phonemes) as represented in the data. The data of the five male and the five female voices are analyzed as two separate samples and the results are compared with those obtained using classical discriminant analysis.

1 Sources of variability in the speech signal

It has been widely recognized, at least since the early 'visible speech' era in the late 1940s (Kopp & Green 1946), that the speech signal is affected by an intimate interaction of several variability sources. This interaction was the major cause of the near-failure of teaching 'speech reading' to the deaf, initially one of the most promising applications of 'visible speech'. If the speech units to be recognized by the deaf – or indeed, any 'speech recognizer' (man or machine) are phoneme-sized – it may be useful to distinguish the following sources of variability:

(i) systematic
  (1) intrinsic
    (a) linguistic
    (b) paralinguistic
  (2) extrinsic
(ii) unsystematic (random).

DOI:10.1017/S002510030100202X Journal of the International Phonetic Association (2001) 31/2

A full list is not intended here, nor is it certain that all SYSTEMATIC sources of variability have been identified, but a tentative list may look something like this: (1a) articulatory, coarticulatory, phonemic, tempo-related, style-related; (1b) speaker's natural voice features, emotions, speaker's incidental conditions (such as a temporary hoarseness or inebriation); (2) frequency characteristics of the channel, noise, reverberation, cross-talk, etc.

The interaction of the different sources of variability has been a serious problem in Automatic Speech Recognition (ASR), and has lead to a situation, not palatable to linguists and phoneticians, where contemporary ASR devices depend on the stochastic nature of the speech signal rather than on phonetic features as was the case in the early attempts at ASR (Denes 1975, Schroeder 1985, Ainsworth 1988, Holmes 1988, De Mori 1998, Keller 1999). Of the main aspects named above, extrinsic (environmental) problems belong to engineering (technological) acoustics. The intrinsic ones come within the area of ACOUSTIC PHONETICS.

When quantitative experimental data are analyzed statistically, it may be impracticable, or even impossible to include many, much less all, sources of variability as such, not only because it may not be known how to define the number of classes to be distinguished, but also because a relatively large number of variables may require a prohibitively large sample size. Also, an excessively complex statistical model may be inoperative. So, speech data may make it necessary to ignore or exclude some variability sources. This is why early applications of even simple statistical models to acoustic speech data required drastic simplifications. For instance, many early works on vowels based on spectrum analysis were performed on isolated steady-state utterances spoken by a relatively small number of talkers (e.g. Arnold et al. 1958, Jassem, Krzyśko & Dyczkowski (1972), Papçun 1980), or vowels spoken in phonologically identical context (e.g. Peterson & Barney 1952). This is permissible if there are substantive reasons to assume that the sample is sufficiently representative for the particular analysis. Also, in parametric statistical analysis, it is desirable that the distribution of the independent variable(s) should not differ very significantly from uni- or multinormal.

It is a hackneyed phrase of NON-MATHEMATICAL (qualitative) general phonetics that sounds of SPOKEN LANGUAGE differ infinitely. This position often does not clearly distinguish between SYSTEMATIC and RANDOM variability and is therefore largely non-productive. For both theoretical-linguistic and applicational purposes, a STATISTICAL approach is much more fruitful, and in fact is the best way out of the ineffective position of quasi-unordered 'infinite variability', which makes any empirical classification and taxonomy either impossible or untenable (ad hoc, heuristic or arbitrary). When two variability sources are represented in one set of data, as in the present study, the effect of one variability source may be reduced by NORMALIZATION.

Several normalization procedures have been proposed over the last thirty-five years or so in the analysis of vowel spectra. Their purpose has been to find a mathematical/geometrical formula that minimizes inter-speaker variability whilst maximizing phonemic differences. The data was, in most cases, the values of the first three formant frequencies though other representations of the spectral properties of phones (mostly vowels) have also been used, such as the levels in neighbouring frequency bands, e.g. by Papçun (1980) or Pols (1977). The best known normalization methods have been compared, and assessed, by Deterding (1990), who obtained average scores of around 90 percent on relatively small samples (5 cases of each phoneme per speaker) for the English MONOPHTHONGS in the context /hVd/ using the various normalization algorithms.

One specific source of variability may be featured by one variable, or a set of several variables. For instance, phonemic distinctions between vowels require a set of variables such as F_1, F_2 and F_3. On the other hand, the data may be such that, say, two

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

different sources of variability are described in terms of THE SAME VARIABLE(S), as in the present study, where we consider five variables, F_1, F_2, F_3, F_4 and F_5 – the vowel formant frequencies – and two systematic variability sources: phonemic and speaker specific. All extrinsic variation has been eliminated by recording the material in a sound-treated room using professional (digital) recording equipment. All variability related to connected speech was ignored as the material consisted of isolated words. Also, variation due to intonation was eliminated by having the speakers utter the words on a monotone. It was assumed that coarticulatory effects could be treated as random variation due to the quasi-random selection of the words. The interaction between the two systematic variability sources was investigated in Krzyśko, Jassem & Wołyński (1999). The present study is specifically concerned with the PHONEMIC distinctions between Polish vowels, whilst INTER-SPEAKER VARIABILITY IS REDUCED BY STATISTICAL NORMALIZATION. For the same data, the effect of normalization over the PHONEMES for a discrimination of the SPEAKERS is treated in Jassem, Krzyśko & Wołyński (2000), and a discriminant analysis (linear discriminant functions) of three variability sources: phonemic, speaker gender and speaker identity is presented in Jassem (1999 & 2000).

2 The data

A special word-list was composed, with the principal aim being that the distribution of the individual Polish phonemes (vowels as well as consonants) should be as nearly uniform as possible (Jassem 1997). This ensured that the a priori probabilities of each of the six vowel phonemes (classes, or groups, in the statistical sense) were nearly equal. The list was read by five male and five female voices with no speech impairment. Conventional digital spectrograms were made of each reading of each word, and the moments of minimal formant movements (the 'targets') were found by eye. At 10 ms intervals around the targets, three spectral sections (A, B, C) were made and the instantaneous frequencies of formants one to five were found using 14– or 16–order LPC within a frequency range safely above F_5. A total of 3390 spectra each represented by the five formant frequencies were first STATISTICALLY NORMALIZED so as to minimize inter-speaker variability. This normalization consisted in computing standard deviations separately for each variable within each of the two groups of speakers, male and female, and dividing each original value of the variables by the appropriate SD.

3 The statistical model

A vowel spectrum is described here in terms of five normalized formant frequencies $F_1^*, F_2^*, F_3^*, F_4^*, F_5^*$. Let $\mathbf{F}^* = (F_1^*, \ldots, F_5^*)'$ denote a random vector whose elements are the five normalized formant frequencies. We shall assume that for each vowel spectrum of a vowel type (a phoneme), the vector $\mathbf{F}^*$ has a five-dimensional normal distribution, with an expected value $\boldsymbol{\mu}$ and positively defined covariance matrix $\boldsymbol{\Sigma}$. The covariance matrices relating to the respective vowel types are not equal. We adopt a two-step classification procedure. In the first step we consider all possible pairs of vowel types (groups) and estimate the probability of each spectrum under classification belonging to one and to the other population forming the pairs. In the second step, the probabilities computed for the individual pairs will be combined, and we shall compute the probabilities that each individual spectrum represents one of the six populations.

In order to classify the individual pairs of spectra, we shall use the linear

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

discriminant function as shown in Krzyśko (1999). That function was constructed for the classification of only two populations and under the assumption that the combined probability distribution of the observed p features of the object to be classified is a multidimensional normal distribution, while the vectors of the expected values and the respective covariance matrices in the two populations are not equal. We shall now describe this function.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)'$ be two independent random vectors such that $\mathbf{X} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{Y} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive definite.

For each $\mathbf{a} \in \mathbf{R}^p$, $\mathbf{a} \neq \mathbf{0}$, and each $c \in \mathbf{R}$ let $\mathrm{R}(\mathbf{a}'\mathbf{x}, c)$ denote the discriminant rule that assigns the observation $\mathbf{x}$ to the population π if $\mathbf{a}'\mathbf{x} \leq c$ and to the population π_2 if $\mathbf{a}'\mathbf{x} > c$.

Let us observe that

$$\mathbf{a}'\mathbf{X} \sim N_p(\mathbf{a}'\boldsymbol{\mu}_1, \mathbf{a}'\boldsymbol{\Sigma}_1\mathbf{a}) \text{ and } \mathbf{a}'\mathbf{Y} \sim N_p(\mathbf{a}'\boldsymbol{\mu}_2, \mathbf{a}'\boldsymbol{\Sigma}_2\mathbf{a}).$$

We have

$$(1)\quad \mathrm{F}_1(c) = \mathrm{P}(\mathbf{a}'\mathbf{X} \leq c) = \mathbf{P}\left(\frac{\mathbf{a}'\mathbf{X} - \mathbf{a}'\boldsymbol{\mu}_1}{(\mathbf{a}'\boldsymbol{\Sigma}_1\mathbf{a})^{1/2}} \leq \frac{c - \mathbf{a}'\boldsymbol{\mu}_1}{(\mathbf{a}'\boldsymbol{\Sigma}_1\mathbf{a})^{1/2}}\right) = \boldsymbol{\Phi}\left(\frac{c - \mathbf{a}'\boldsymbol{\mu}_1}{(\mathbf{a}'\boldsymbol{\Sigma}_1\mathbf{a})^{1/2}}\right)$$

and

$$(2)\quad \mathrm{F}_2(c) = \mathrm{P}(\mathbf{a}'\mathbf{Y} \leq c) = \boldsymbol{\Phi}\left(\frac{c - \mathbf{a}'\boldsymbol{\mu}_2}{(\mathbf{a}'\boldsymbol{\Sigma}_2\mathbf{a})^{1/2}}\right),$$

where $\boldsymbol{\Phi}$ is the c.d.f. of an $N(0,1)$ random variable.

Each discriminant rule is characterized in terms of the two probabilities of misclassification or in terms of the two conditional probabilities of correct classification. The probability of misclassifying an observation when it comes from the first population is

$$\mathrm{P}(\pi_2 \mid \pi_1) = \mathrm{P}(\mathbf{a}'\mathbf{X} > c) = 1 - \mathrm{P}(\mathbf{a}'\mathbf{X} \leq c) = 1 - \mathrm{F}_1(c)$$

and the probability of misclassifying an observation when it comes from the other population is

$$\mathrm{P}(\pi_1 \mid \pi_2) = \mathrm{P}(\mathbf{a}'\mathbf{Y} \leq c) = \mathrm{F}_2(c),$$

where $\mathrm{F}_1(c)$ and $\mathrm{F}_2(c)$ are given by (1) and (2), respectively.

The corresponding conditional probabilities of correct classification are

$$\mathrm{P}(\pi_1 \mid \pi_1) = \mathrm{P}(\mathbf{a}'\mathbf{X} \leq c) = \mathrm{F}_1(c)$$

and

$$\mathrm{P}(\pi_2 \mid \pi_2) = \mathrm{P}(\mathbf{a}'\mathbf{Y} \leq c) = \mathrm{F}_2(c)$$

The probability $\mathrm{P}(\pi_1 \mid \pi_1)$ is called specificity of the discriminant rule and the probability $\mathrm{P}(\pi_2 \mid \pi_2)$ is called sensitivity of the discriminant rule.

In a parametric representation, the curve of the form

$$x = \mathrm{F}_1(c),\ y = 1 - \mathrm{F}_2(c),\ -\infty \leq c \leq \infty,$$

is called the Relative Operating Characteristic (ROC) curve of the class rules $\mathrm{R}(\mathbf{a}'\mathbf{x}, \cdot)$.

An ROC curve inevitably passes (1,0) by selecting a large value of c and (0,1) by selecting a low value of c and is concave.

Some investigators plot

$$x = 1 - \mathrm{F}_1(c),\ y = 1 - \mathrm{F}_2(c),\ -\infty \leq c \leq \infty,$$

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

instead and use the term receiver operating characteristic curve (Swets & Pickett 1982).

The area D(**a**) under the ROC curve is the index which evaluates the accuracy of a class of the discriminant rules R(**a**′**x**,*c*). A large area indicates that the linear combination **a**′**x** discriminates well between the two populations being compared. This index has a simple probabilistic interpretation (Krzyśko 1999). The area D(**a**) under the ROC curve is equal to the probability that the random variable **a**′**Y** is stochastically large than the random variable **a**′**X**:

$$D(\mathbf{a}) = P(\mathbf{a}'\mathbf{Y} > \mathbf{a}'\mathbf{X}),$$

where **X** and **Y** are two independent random vectors such that $\mathbf{X} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{Y} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $\boldsymbol{\Sigma}_1 > 0$, $\boldsymbol{\Sigma}_2 > 0$, $\mathbf{a} \in \mathbf{R}^p$, $\mathbf{a} \neq \mathbf{0}$.

A comparison of the areas under the different ROC curves may be used to determine which linear combination **a**′**X** is best.

One discriminant function is better than another if each probability of misclassification of the former is not greater than the corresponding one of the latter and at least one is less. A discriminant function is admissible if there is no other function that is better.

The linear discriminant function for which the area under the corresponding ROC curve is maximized and which is admissible within the class of linear rules has the following form (Krzyśko 1999):

$$u(\mathbf{x}) = \mathbf{a}'\mathbf{x} - c, \tag{3}$$

where

$$\mathbf{a} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \tag{4}$$

and

$$c = \mathbf{a}'\boldsymbol{\mu}_1 + \mathbf{a}'\boldsymbol{\Sigma}_1\mathbf{a} = \mathbf{a}'\boldsymbol{\mu}_2 - \mathbf{a}'\boldsymbol{\Sigma}_2\mathbf{a}. \tag{5}$$

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, then

$$\mathbf{a}'\mathbf{x} - c = \tfrac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\,\boldsymbol{\Sigma}^{-1}[\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)].$$

We see that $\mathbf{a}'\mathbf{x} - c$ is a well-known Fisher linear discriminant function.

Note also that for **a** of the form (4) the maximum area D(**a**) is equal to

$$D(\mathbf{a}) = \Phi\,[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\,(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)]^{1/2}. \tag{6}$$

But $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\,(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \geq 0$ and $\Phi\,[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\,(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)]^{1/2} \geq \tfrac{1}{2}$. Hence

$$\tfrac{1}{2} \leq D(\mathbf{a}) \leq 1,$$

with D close to 1 indicating that the *p* characteristics distinguish well between the population π_1 and π_2, and D close to $\frac{1}{2}$ indicating that the two populations are not well separated.

Let us recall that if $\mathbf{a}'\mathbf{x} - c \leq 0$, then we assign the observation **x** to population π_1 and if $\mathbf{a}'\mathbf{x} - c > 0$, the observation **x** is assigned to population π_2.

The conditional probability of observation **x** belonging to population π_1 under the condition that this observation belongs to π_1 or $\boldsymbol{\pi}_2$ is defined as

$$r_{12}(\mathbf{x}) = P(\pi_1 \mid \pi_1 \text{ or } \pi_2) = \frac{1}{1 + \exp(\mathbf{a}'\mathbf{x} - c)}. \tag{7}$$

The conditional probability of observation **x** belonging to population π_2 under the condition that observation **x** belongs to population π_1 or $\boldsymbol{\pi}_2$ is defined as

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

$$r_{21}(\mathbf{x}) = P(\pi_2 \mid \pi_2 \text{ or } \boldsymbol{\pi}_1) = \frac{\exp(\mathbf{a}'\mathbf{x} - c)}{1 + \exp(\mathbf{a}'\mathbf{x} - c)}. \tag{8}$$

Clearly,

$$r_{12}(\mathbf{x}) + r_{21}(\mathbf{x}) = 1.$$

The inequality $r_{12}(\mathbf{x}) \geq r_{21}(\mathbf{x})$ indicates that $\mathbf{x} \in \pi_1$. But this inequality is equivalent to inequality $\mathbf{a}'\mathbf{x} - c \leq 0$.

If we have k populations $\pi_1, \ldots, \pi_k$, then for each pair (π_i, π_j), $i \neq j$, we can compute $r_{ij}(\mathbf{x}) = P(\pi_i \mid \pi_i \text{ or } \pi_j)$.

Hastie & Tibshirani (1998) suggested that the decision on the assignment of observation $\mathbf{x}$ to one of the k populations should be based on the value of the following function

$$\bar{p}_i(\mathbf{x}) = 2\frac{\sum_{j \neq i}^{k} r_{ij}(\mathbf{x})}{k \quad k(k-1)}, \; i = 1,2, \ldots, k. \tag{9}$$

This function takes into account all the conditional probabilities $r_{ij}(\mathbf{x})$ after the populations have been joined into pairs.

We assign the observation $\mathbf{x}$ to that population which has the highest value of the function $\bar{p}_i(\mathbf{x})$.

Let us denote the a posteriori probability of the observation $\mathbf{x}$ belonging to population π_i, $i = 1, \text{K}, k$ by $p_i(\mathbf{x})$. Hastie & Tibshirani (1998) presented an algorithm for estimating the probability $p_i(\mathbf{x})$ based on the values $r_{ij}(\mathbf{x})$. If $\hat{p}_i(\mathbf{x})$ is an estimator of the probability $p_i(\mathbf{x})$, then

$$\bar{p}_i(\mathbf{x}) > \bar{p}_j(\mathbf{x}) \text{ if and only if } \hat{p}_i(\mathbf{x}) > \hat{p}_j(\mathbf{x})$$

Likewise,

$$\bar{p}_i(\mathbf{x}) = \bar{p}_j(\mathbf{x}) \text{ if and only if } \hat{p}_i(\mathbf{x}) = \hat{p}_j(\mathbf{x})$$

These relations make it possible to classify the observation $\mathbf{x}$ based on the values of the functions $\bar{p}_1(\mathbf{x}), \bar{p}_2(\mathbf{x}), \ldots, \hat{p}_k(\mathbf{x})$.

Clearly, using the algorithm given by Hastie & Tibshirani (1998) it is not only possible to assign the observation $\mathbf{x}$ to just one of the k populations $\pi_1, \ldots, \pi_k$, but also to compute the estimated posterior probability $\hat{p}_i(\mathbf{x})$ that $\mathbf{x}$ belongs to π_i, $i = 1, \ldots, k$.

In our case, we have six populations corresponding to the six Polish vowel phonemes /i ɨ e a o u/. The characteristic features of the vowel spectra are the NORMALIZED formant frequencies $F1^*$, K, $F5^*$. The vowel spectra were identified separately for the female and the male speakers. Table 1 presents the sample sizes for the five female voices. The same figures describe the sample sizes for the male voices.

Table 1 Number of the individual phoneme realizations for each speaker gender.

i	ɨ	e	a	o	u	total
270	270	300	285	300	270	1695

For the individual formant frequencies, the standard deviations were computed separately for the female and the male voices and, subsequently, all values of formant frequencies were divided by these standard deviations. This is (as mentioned above) how the normalized formant frequencies were obtained. The means of the normalized formant frequencies, separately for the male and the female voices, are contained in tables 2 and 3. Tables 4 and 5 contain the sample covariance matrices for the individual phonemes, separately for the male and the female voices.

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

Table 2 Normalized means; female voices.

	F_1^*	F_2^*	F_3^*	F_4^*	F_5^*
i	6.689	16.430	13.557	15.643	17.359
ɨ	7.712	10.308	17.515	16.784	16.686
e	5.509	7.193	14.944	20.269	19.078
a	4.790	10.425	8.670	11.532	13.950
o	4.580	10.659	10.419	13.898	15.049
u	8.513	6.184	12.772	12.015	12.603

Table 3 Normalized means; male voices.

	F_1^*	F_2^*	F_3^*	F_4^*	F_5^*
i	7.524	10.288	9.889	12.567	13.634
ɨ	7.240	10.292	10.115	12.102	13.005
e	4.694	9.491	8.665	13.017	12.274
a	6.521	10.855	11.170	15.741	13.463
o	5.359	11.715	6.371	14.920	13.038
u	7.499	6.768	9.042	10.654	15.726

Table 4 Total correlation matrix for F_1^*, F_2^*, F_3^*,F_4^*, F_5^*; female voices.

		F_1^*	F_2^*	F_3^*	F_4^*	F_5^*
F_1*	i	1	0.154	−0.173	0.160	−0.122
	ɨ		−0.149	0.133	0.402	0.436
	e		−0.436	−0.311	0.158	0.427
	a		−0.122	0.045	−0.162	0.209
	o		0.367	0.119	0.170	0.240
	u		0.273	0.088	0.203	−0.079
F_2*	i	0.154	1	0.102	0.166	0.239
	ɨ	−0.149		0.202	0.054	−0.068
	e	−0.436		0.583	0.002	−0.112
	a	−0.122		0.266	0.339	0.130
	o	0.367		0.027	0.130	0.115
	u	0.273		−0.001	−0.041	−0.044
F_3*	i	−0.173	0.102	1	0.245	0.439
	ɨ	0.133	0.202		0.067	−0.009
	e	−0.311	0.583		0.055	−0.188
	a	0.045	0.266		0.414	0.270
	o	0.119	0.027		−0.022	0.039
	u	0.088	−0.001		0.289	0.277
F_4*	i	0.160	0.166	0.245	1	0.626
	ɨ	0.402	0.054	0.067		0.713
	e	0.158	0.002	0.055		0.427
	a	−0.162	0.339	0.414		0.656
	o	0.170	0.130	−0.022		0.722
	u	0.203	−0.041	0.289		0.447
F_5*	i	−0.122	0.239	0.439	0.626	1
	ɨ	0.436	−0.068	−0.009	0.713	
	e	0.427	−0.112	−0.188	0.427	
	a	0.209	0.130	0.270	0.656	
	o	0.240	0.115	0.039	0.722	
	u	−0.079	−0.044	0.277	0.447	

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

Table 5 Total correlation matrix for $F_1^*, F_2^*, F_3^*, F_4^*, F_5^*$; male voices.

		F_1^*	F_2^*	F_3^*	F_4^*	F_5^*
F_1^*	i	1	0.129	0.040	−0.066	0.408
	ɨ		0.426	0.462	0.158	0.656
	e		−0.079	0.240	0.024	0.753
	a		0.316	0.473	0.383	0.472
	o		−0.092	0.462	0.406	0.528
	u		0.101	0.411	0.243	0.448
F_2^*	i	0.129	1	0.759	0.777	0.283
	ɨ	0.426		0.806	0.584	0.535
	e	−0.079		0.720	0.424	0.119
	a	0.316		0.256	0.190	0.098
	o	−0.092		−0.072	−0.134	−0.142
	u	0.010		−0.185	−0.237	−0.178
F_3^*	i	0.040	0.759	1	0.602	0.282
	ɨ	0.462	0.806		0.515	0.522
	e	0.240	0.720		0.607	0.386
	a	0.473	0.256		0.491	0.412
	o	0.462	−0.072		0.564	0.349
	u	0.411	−0.185		0.432	0.499
F_4^*	i	−0.066	0.777	0.602	1	0.135
	ɨ	0.158	0.584	0.515		0.306
	e	0.024	0.424	0.607		0.181
	a	0.383	0.190	0.491		0.483
	o	0.406	−0.134	0.564		0.513
	u	0.243	−0.237	0.432		0.539
F_5^*	i	0.408	0.286	0.282	0.135	1
	ɨ	0.656	0.535	0.522	0.306	
	e	0.753	0.119	0.386	0.181	
	a	0.472	0.098	0.412	0.483	
	o	0.528	−0.142	0.349	0.513	
	u	0.448	−0.178	0.499	0.539	

Note that the figures in tables 2 and 3 are abstract numbers, and being derived from the original variables and their deviations, they do not directly map the formant frequencies.

Next, the $\binom{6}{2} = 15$ pairs of vowel phonemes were considered. For each pair, a linear discriminant function was created $u(\mathbf{F}^*) = \mathbf{a}'\mathbf{F} - c$ replacing the vectors of means and the covariance matrices by their respective estimators from the samples. In a given pair, a change of the numbered populations only results in the change of the sign in the functions. Next, for each pair (π_i, π_j), $i, j = 1, \ldots, 6$, $i \neq j$, the values of $r_{ij}(\mathbf{F}^*)$ were computed, as well as – in agreement with (9) – the values

$$\bar{p}_i(\mathbf{F}^*) = \tfrac{1}{2}\sum_{j\neq i}^{6} \mathbf{r}_{ij}(\mathbf{F}^*), \quad i = 1,\ldots,6.$$

On the basis of the functions $\bar{p}_i(\mathbf{F}^*)$ the 1695 spectra were classified with respect to the six phonemes, separately for the female and for the male voices.

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

4 Results

4.1 The predictability of the models with respect to the three sets of data

As stated above, the measurements of the spectral features of each the vowels were taken (A) at the 'target', (B) at −10 ms off the target and at +10 ms off the target. We were first interested to test the predictability of the statistical models with respect to the three sets of data. The results of an analysis of the model presented in section 3 above are shown in tables 6 and 9. For comparison, the analogous results of two standard classification procedures, with linear and quadratic discriminant functions, are presented in tables 7–8 and 10–11.

Each of the three models was tested using three methods:

1 Resubstitution. The cases of the training set are used to find the allocation rule and are then substituted into it to estimate its performance.
2 Partitioning (a). Measurements A and B were used as the training set, and measurements C as the test set. The cases in the test set were classified using the standard deviations as estimated in the training set.
3 Partitioning (b). The partitioning of the sample was the same as in 2 above, but the standard deviations were now estimated from the test set.

Table 6 Confusion matrix: modified discriminant analysis; female voices.

	Method	i	ɨ	e	a	o	u	%
i	**1**	270	0	0	0	0	0	100
	2	90	0	0	0	0	0	100
	3	90	0	0	0	0	0	100
ɨ	**1**	0	270	0	0	0	0	100
	2	0	90	0	0	0	0	100
	3	0	90	0	0	0	0	100
e	**1**	0	0	300	0	0	0	100
	2	0	0	100	0	0	0	100
	3	0	0	100	0	0	0	100
a	**1**	0	0	0	272	13	0	95.44
	2	0	0	0	93	2	0	97.89
	3	0	0	0	79	16	0	83.16
o	**1**	0	0	0	29	271	0	90.33
	2	0	0	0	8	92	0	92.00
	3	0	0	0	20	80	0	80.00
u	**1**	0	0	0	0	0	270	100
	2	0	0	0	0	0	90	100
	3	0	0	0	0	1	89	98.89
						Total	**1**	97.52
							2	98.23
							3	93.45

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

Table 7 Confusion matrix: linear discriminant analysis; female voices.

	Method	i	ɨ	e	a	o	u	%
i	**1**	270	0	0	0	0	0	100
	2	90	0	0	0	0	0	100
	3	90	0	0	0	0	0	100
ɨ	**1**	0	270	0	0	0	0	100
	2	0	90	0	0	0	0	100
	3	0	89	0	0	0	0	98.89
e	**1**	0	0	300	0	0	0	100
	2	0	0	100	0	0	0	100
	3	0	0	100	0	0	0	100
a	**1**	0	0	0	247	38	0	86.67
	2	0	0	0	89	6	0	93.68
	3	0	0	0	78	17	0	82.11
o	**1**	0	0	0	19	281	0	93.67
	2	0	0	0	5	95	0	95.00
	3	0	0	0	15	85	0	85.00
u	**1**	0	0	0	0	0	270	100
	2	0	0	0	0	0	90	100
	3	0	0	0	0	1	90	100
						Total	**1**	96.64
							2	98.05
							3	94.16

Table 8 Confusion matrix: quadratic discriminant classification; female voices.

	Method	i	ɨ	e	a	o	u	%
i	**1**	270	0	0	0	0	0	100
	2	90	0	0	0	0	0	100
	3	90	0	0	0	0	0	100
ɨ	**1**	0	270	0	0	0	0	100
	2	0	90	0	0	0	0	100
	3	0	90	0	0	0	0	100
e	**1**	0	0	300	0	0	0	100
	2	0	0	100	0	0	0	100
	3	0	0	100	0	0	0	100
a	**1**	0	0	0	263	22	0	92.28
	2	0	0	0	91	4	0	95.79
	3	0	0	0	79	16	0	83.16
o	**1**	0	0	0	27	273	0	91.00
	2	0	0	0	5	95	0	95.00
	3	0	0	0	16	84	0	8400
u	**1**	0	0	0	0	0	270	100
	2	0	0	0	0	0	90	100
	3	0	0	0	0	1	89	98.89
						Total	**1**	97.11
							2	98.41
							3	94.16

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

Table 9 Confusion matrix: modified discriminant analysis; male voices.

	Method	i	ɨ	e	a	o	u	%
i	**1**	194	69	1	3	0	3	71.85
	2	57	24	4	5	0	0	63.33
	3	49	34	3	1	1	2	54.44
ɨ	**1**	58	207	5	0	0	0	76.67
	2	19	68	3	0	0	0	75.56
	3	44	40	6	0	0	0	44.44
e	**1**	4	7	245	43	1	0	81.67
	2	0	3	81	15	1	0	81.00
	3	0	3	83	13	1	0	83.00
a	**1**	8	4	5	267	1	0	93.68
	2	1	3	3	88	0	0	92.63
	3	0	0	1	88	6	0	92.63
o	**1**	0	0	0	7	293	0	97.67
	2	0	0	0	1	99	0	99.00
	3	0	0	0	1	99	0	99.00
u	**1**	4	0	0	0	0	266	98.52
	2	3	0	1	0	0	86	95.56
	3	0	0	0	0	0	90	100
						Total	**1**	86.84
							2	84.78
							3	79.47

Table 10 Confusion matrix: linear discriminant analysis; male voices.

	Method	i	ɨ	e	a	o	u	%
i	**1**	193	65	1	8	0	3	71.48
	2	53	23	4	9	0	1	58.89
	3	59	17	4	1	0	9	65.56
ɨ	**1**	75	187	8	0	0	0	69.26
	2	20	67	3	0	0	0	74.44
	3	42	40	7	0	0	1	44.44
e	**1**	18	1	244	37	0	0	81.33
	2	1	0	86	13	0	0	86.00
	3	1	0	88	11	0	0	88.00
a	**1**	6	4	9	266	0	0	93.33
	2	2	1	5	87	0	0	91.58
	3	0	0	1	91	3	0	95.79
o	**1**	0	0	0	8	292	0	97.33
	2	0	0	0	1	99	0	99.00
	3	0	0	0	1	99	0	99.00
u	**1**	0	0	0	0	0	270	100
	2	0	0	0	0	0	90	100
	3	0	0	0	0	0	90	100
						Total	**1**	85.66
							2	85.31
							3	82.65

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

Table 11 Confusion matrix for quadratic discriminant analysis; male voices.

	Method	i	ɨ	e	a	o	u	%
	1	185	74	9	0	0	2	68.52
i	**2**	63	21	4	2	0	0	70.00
	3	66	17	3	1	0	3	73.33
	1	56	204	10	0	0	0	75.56
ɨ	**2**	18	67	5	0	0	0	74.44
	3	41	42	7	0	0	0	46.67
	1	3	7	258	32	0	0	86.00
e	**2**	0	2	88	10	0	0	88.00
	3	0	2	90	8	0	0	90.00
	1	12	1	7	264	1	0	92.63
a	**2**	4	0	2	89	0	0	93.68
	3	0	0	1	92	2	0	96.84
	1	0	0	0	8	292	0	97.33
o	**2**	0	0	0	1	99	0	99.00
	3	0	0	0	2	98	0	98.00
	1	3	0	0	0	0	267	98.89
u	**2**	0	0	0	0	0	90	100
	3	0	0	0	0	1	90	100
						Total	**1**	86.73
							2	87.79
							3	84.60

4.2 Summary

The results can be summed up under 5 headings: (i) Speaker gender, (ii) Vowel phonemes (vowel-gender interaction), (iii) Testing methods, (iv) Analysis methods, and (v) Predictivity.

(i) All results are better in the female voices than the male voices.
(ii) For the front vowels /i ɨ e/, the results for the female voices are better than for the male voices everywhere.
(iii) There are no significant differences between the testing methods (ii) and (iii).
(iv) The quadratic discriminant functions tend to result in better scores than the linear functions, but this difference is not significant. The new method proposed in this article (cf. section 3 above) results in scores that are, on the whole, closer to the scores obtained with quadratic discriminant functions than those obtained with the linear discriminant functions.
(v) The overall testing result is that the models are highly predictive with respect to the three sets of measurement data.

4.3 The inter-speaker predictability of the three models

Another aspect of our data is its PREDICTABILITY with respect to vowel classification for the individual voices. This reduces to the question: how well can the vowels of each of the speakers in a set (male or female, in our case) be predicted on the basis of the data from the remaining speakers in the set? We have tested all the ten of our speakers in this way, each one against the remaining four in the set. Detailed results will be given here, in tables 12 and 13, for one of the male and one of the female voices, respectively. For reasons of space, the results for the other eight speakers will be treated summarily.

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

Table 12 Partitioning test for speaker JI (male).

JI

	Method	i	ɨ	e	a	o	u	%
i	**1**	25	29	0	0	0	0	46.30
	2	20	34	0	0	0	0	37.04
	3	24	30	0	0	0	0	44.44
ɨ	**1**	7	47	0	0	0	0	87.04
	2	10	44	0	0	0	0	81.48
	3	7	47	0	0	0	0	87.04
e	**1**	3	2	55	0	0	0	91.67
	2	5	6	49	0	0	0	81.67
	3	13	2	45	0	0	0	75.00
a	**1**	1	0	16	40	0	0	70.18
	2	1	0	16	40	0	0	70.18
	3	1	0	22	34	0	0	59.65
o	**1**	0	0	0	0	60	0	100
	2	0	0	0	0	60	0	100
	3	0	0	0	0	60	0	100
u	**1**	0	0	0	0	0	54	100
	2	0	0	0	0	0	54	100
	3	0	0	0	0	0	54	100
						Total	**1**	82.8
							2	78.76
							3	77.88

Table 13 Partitioning test for speaker AI (female).

AI

	Model	i	ɨ	e	a	o	u	%
i	**LDF**	54	0	0	0	0	0	100
	QDF	54	0	0	0	0	0	100
	MDF	54	0	0	0	0	0	100
ɨ	**LDF**	0	54	0	0	0	0	100
	QDF	0	52	0	0	0	2	96.30
	MDF	0	52	0	0	0	2	96.30
e	**LDF**	0	0	60	0	0	0	100
	QDF	0	0	60	0	0	0	100
	MDF	0	0	60	0	0	0	100
a	**LDF**	0	0	0	57	0	0	100
	QDF	0	0	0	57	0	0	100
	MDF	0	0	0	57	0	0	100
o	**LDF**	0	0	0	5	55	0	91.67
	QDF	0	0	0	8	52	0	86.67
	MDF	0	0	0	5	55	0	91.67
u	**LDF**	0	0	0	0	0	54	100
	QDF	0	0	0	0	0	54	100
	MDF	0	0	0	0	0	54	100
						Total	**1**	98.53
							2	97.05
							3	97.94

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

Table 14 Mean percent scores for all vowels by speaker.

speaker	JK	PW	TZ	WJ	AD	AL	KK	LR
score	67.85	78.76.	81.12	75.52	100	99.71	92.04	91.45

JK, PW, TZ and WJ are male speakers.
AD, AL, KK and LR are female speakers.

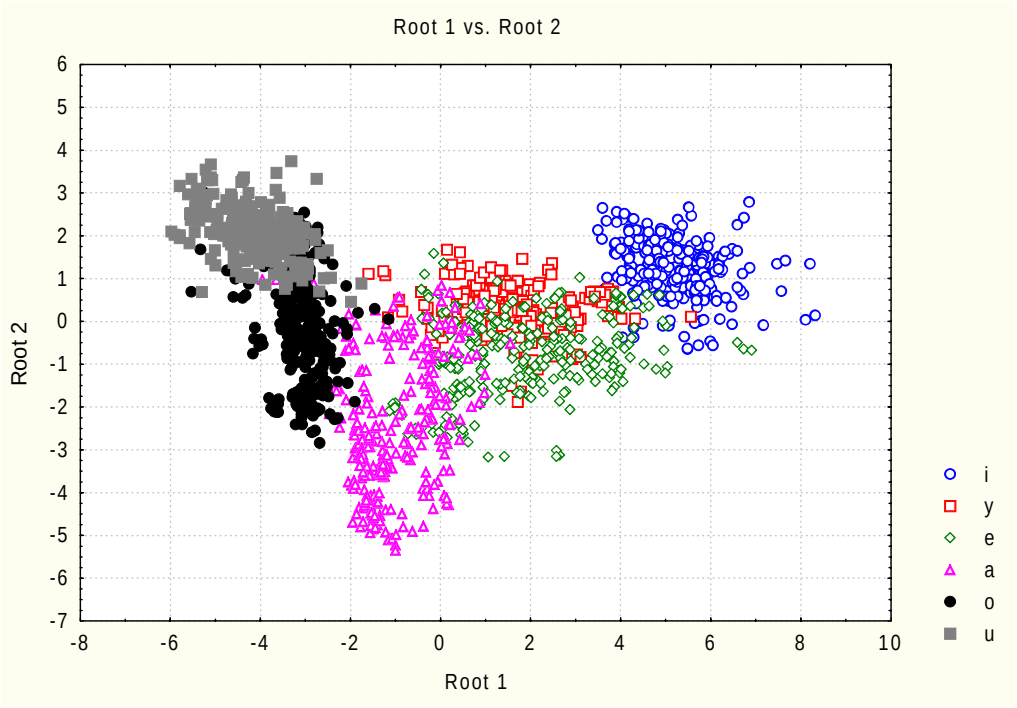

Figure 1 Unnormalized female vowels in the (Root 1, Root 2) plane.

The remaining results, for the model presented in section 3 above, are shown in table 14 in a condensed form.

A (partial) visual representation of the reduction of speaker-related variability and the enhancement of the phonetic distinctiveness of the vowel spectra by the normalization proposed here can be gained from figures 1 and 2. These map our female vowel spectra into a Root 1 vs. Root 2 plane of a classical linear discriminant analysis. The complete images are multi-dimensional but roots no. 1 and 2 are the most strongly distinctive.

5 Conclusions

If the formant frequencies of Polish vowels (all monophthongal) are measured at or near the 'targets', and then very simply normalized statistically, tokens of the six phonemes /i ɨ e a o u/ can be classified, and recognized, with high accuracy, speaker-independently, separately in male and female voices. This hypothesis was confirmed on a large corpus including 3390 cases of vowel spectra in special word lists constructed in such a way that each of the six vowel phonemes occurred in various phonetic contexts.

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

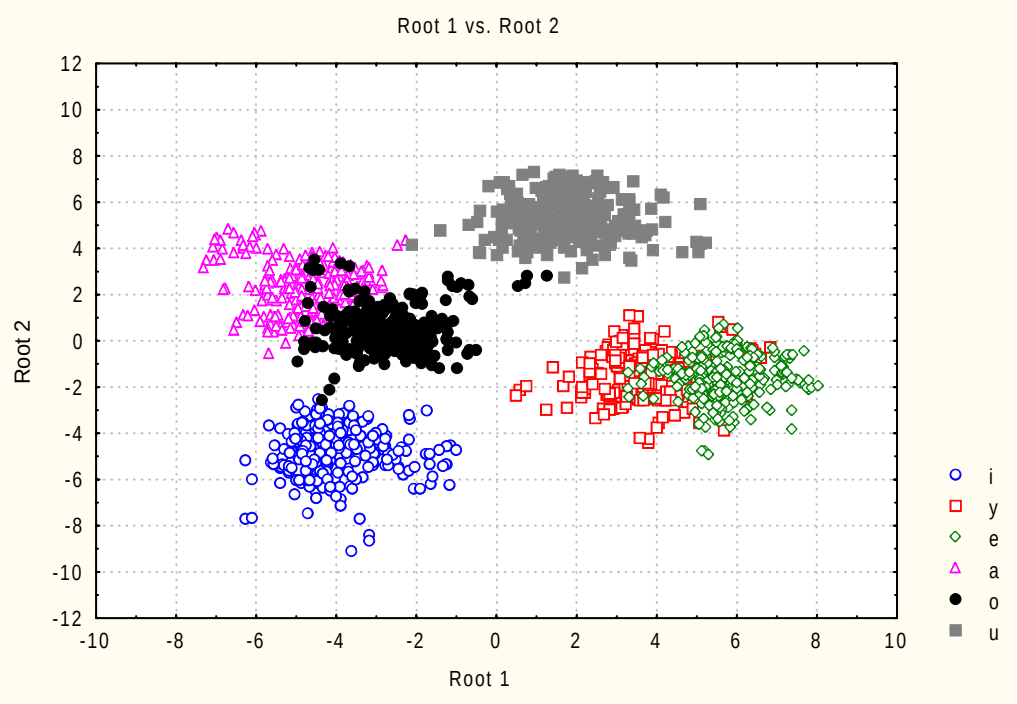

Figure 2 Normalized female vowels in the (Root 1, Root 2) plane.

The lists were read by five male and five female speakers, and the measurements were analyzed using three discriminant models, using substitution and partitioning tests.

Ackowledgement

The present study was financed by the Polish State Committee for Scientific Research (KBN) within the framework of project No. 8 T.11D.029.16 'Variability and invariance of the speech signal and its segmentation'.

References

Ainsworth, W. A. (1988). *Speech Recognition by Machine*. London: P. Peregrinus.

Arnold, G. F., Denes, P., Gimson, A. C., O´connor, J. D. & Trim, L. M. (1958). The synthesis of English vowels. *Language and Speech* **1**, 114–125.

De Mori, R. (ed.) (1998). *Spoken Dialog with Computers*. London: Academic Press.

Denes, P. (1975). Automatic speech recognition. In Reddy, D. R. (ed.), *Speech Recognition*, 73–83. New York: Academic Press.

Deterding, H. (1990). *Speaker Normalization for Automatic Speech Recognition*. Ph.D. dissertation, University of Cambridge.

Fry, D. B. & Denes, P. (1956). Experiments in mechanical speech recogniton. In Cherry, C. (ed.), *Information Theory: Third London Symposium*, 206–212. Butterworth, London.

Hastie, T. & Tibshirani, R. (1998). Classification by pairwise coupling. *The Annals of Statistics* **26**, 451–471.

Holmes, J. N. (1988). *Speech Synthesis and Recognition*.Wokingham: Van Nostrand.

Jassem, W. (1997). Zrównoważone częstotliwościowo i fonetycznie polskie listy wyrazowe [Polish

https://doi.org/10.1017/S002510030100202X Published online by Cambridge University Press

phonetically balanced and frequency-weghted word lists]. In Jassem, W. & Basztura, Cz. (eds.), *Speech and Language Technology* **1**, 71–99, Poznań: Polish Phonetic Association.

Jassem, W. (1999 & 2000). Formants of the Polish vowels as phonemic and speaker-related cues: report on a discriminant analysis. In Jassem et al. (eds.), 191–216. Erratum. In Jassem, W., Basztura, Cz. & Demenko, D. (2000). *Speech and Language Technology* **4**, 127–135. Poznań: Polish Phonetic Association.

Jassem, W., Basztura, Cz., Demenko, D. & Jassem, K. (eds.) (1999). *Speech and Language Technology* **3**. Poznań: Polish Phonetic Association.

Jassem, W., Krzyśko, M. & Dyczkowski, A. (1972). *Klasyfikacja samogłosek polskich na podstawie częstotliwości formantów* [Classification of Polish vowels based on formant frequencies]. Warszawa: Instytut Podstawoych Problemów Techniki.

Jassem, W., Krzyśko, M. & Wołyński, W. (2000). Normalisation of Polish vowel spectra. *The Phonetician* **81**, 23–31.

Keller, E. (1999). *Fundamentals of Speech Synthesis and Speech Recognition*. Chichester: Wiley & Sons.

Kopp. G. A. & Green, H. C. (1946). Basic phonetic principles of visible speech. *Journal of the Acoustical Society of America* **18**, 74–89.

Krzyśko, M. (1999). Linear discriminant functions which maximize the area under the ROC curve. *Discussiones Mathematicae: Algebra and Stochastic Methods* **19**, 335–344.

Krzyśko, M., Jassem, W. & Czajka, S. (1999). The formants of Polish vowels: a multivariate analysis of variance with two factors. In Jassem et al. (eds.), 173–189.

Papçun, G. (1980). *How do different speakes say the same vowels? Discriminant analysis of four imitation dialects* (UCLA Working Papers in Phonetcis 48). Los Angeles: UCLA.

Peterson G.E. & Barney H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* **24**, 175–184.

Pols, L. C. W. (1977). *Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words*. Soesterberg: Institute for Perception, TNO.

Schroeder, M. R. (ed.) (1985). *Speech and Speaker Recognition*. Basel: Arger.

Swets, J. A. & Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press.