euronoise

Complexity of acoustic-production-based models of speech perception

Geoffrey Stewart Morrison

Australian National University, School of Language Studies, Building 110, ACT 0200
Canberra, Australia
geoff.morrison@anu.edu.au

Discriminant analysis models trained on acoustic vowel production data have been found to have significant correlation with listeners' perception. Two regularized discriminant analysis models were trained on monolingual speakers' vowels. One model was trained on North Central Peninsular Spanish vowel tokens, and the other on Western Canadian English vowel tokens. For each language the model which resulted in the lowest cross-validated classification error rate was close to the least complex model possible, i.e., close to linear discriminant analysis using the mean of the variances of the acoustic variables but not using the covariances between variables. In order to make cross-language vowel perception predictions the Spanish model was used to classify English vowel tokens and the English model was used to classify Spanish vowel tokens. Results suggest that monolingual North Central Peninsular Spanish listeners would assimilate most tokens of Western Canadian English /i/ and /ɪ/ to Spanish /i/ and /e/ respectively, and thus for this combination of dialects, Spanish-speaking learners of English would not be expected to have difficulty with the English /i/–/ɪ/ contrast.

1 Introduction

It is assumed that listeners learn to categorize speech sounds on the basis of the statistical distribution of acoustic properties of the speech to which they are exposed [1, 2]. Thus a model of human speech perception can also be constructed on the basis of acoustic production data. Discriminant analysis models trained on speech production data have previously been found to have significant correlation with listeners' perception [3, 4, 5]. The present paper trains regularized discriminant analysis models on acoustic data from monolingual speakers' vowels, and for each model determines the level of model complexity which results in the highest cross-validated correct-classification rate. One model is trained and tested on acoustic data from North Central Peninsular Spanish vowels, and another is trained and tested on acoustic data from Western Canadian English vowels (in each case, only a subset of the vowel inventory was used). The Spanish and English models are then used to make predictions as to how monolingual listeners of each language will perceive the vowels of the other language. Such predictions are useful for understanding the vowel-perception problems which a speaker of one language may face when they begin to learn the other language.

2 Data

Seventeen monolingual speakers of North Central Peninsular Spanish (eight male and nine female) were recruited in Vitoria-Gasteiz, Autonomous Region of the Basque Country, Spain. Nineteen monolingual Western Canadian English speakers (eight male and eleven female) were recruited in Edmonton, Alberta, Canada. They read sentences aloud in response to the written prompts: "La próxima palabra es ____" and "The next word is ____" (the Spanish and English sentences have the same meaning). The prompt words were *BIPA*, *BEPA*, and *BEIPA* in Spanish corresponding to /bipa/, /bepa/, and /beipa/, and *BEEPA*, *BIPPA*, *BAYPA*, and *BEPPA* in English corresponding to /bipə/, /bɪpə/, /bepə/, and /bɛpə/. Each speaker read each sentence ten times in randomized order. Recordings were made at a sampling frequency of 44.1 kHz using a Sennheiser HMD 280 PRO headset and a Roland ED UA-30 USB Audio Interface with a Rolls MP13 preamplifier.

The duration and first- and second-formant tracks (F1 and F2 tracks) of all the vowels were measured. The geometric means of these acoustic properties are shown in Fig. 1.

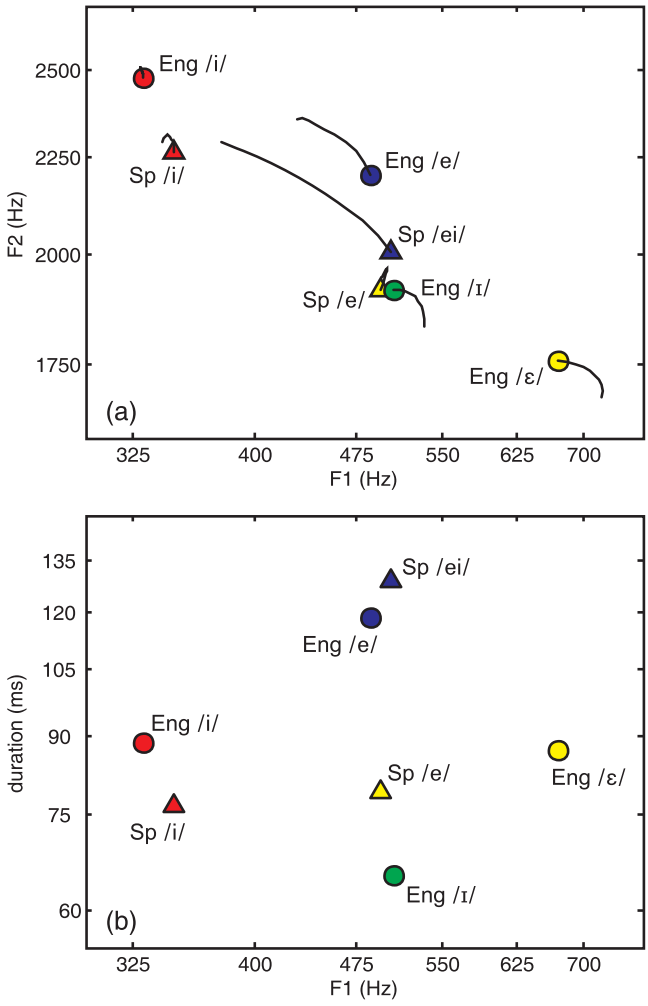

Fig. 1 Gender-balanced geometric means for first and second formants and durations of vowels produced by monolingual Spanish and English speakers. (a) Comet heads represent mean F1 and F2 at 25% of the duration of the vowels and comet tails represent the mean formant trajectories from 25% to 75% of the duration of the vowels. (b) Markers represent mean vowel duration and mean F1 at 25% of the duration of the vowels.

3 Discriminant Analysis Models

Two discriminant analysis models were constructed, one trained on the monolingual Spanish speakers' vowel production data and one trained on the monolingual English speakers' vowel production data. Five acoustic predictor variables were entered into the models: F1 at 25% of the duration of the vowel, the change in F1 from 25% to 75% of the duration of the vowel, F2 at 25% of the duration of the vowel, the change in F2 from 25% to 75% of the duration of the vowel, and duration. Prior to constructing the discriminant analysis models, formant values were normalized using a cross-language version of constant-log-interval normalization [6]. Formant values were kept in the log-scale for entry into the model (the exponents of the means of the normalized formant values are the same as the gender-balanced geometric means in Fig. 1). Duration was independently normalized using the same procedure.

There are two basic variants of discriminant analysis, quadratic which makes use of a separate estimate of the within-group covariance matrix for each category, and linear which makes use of a single pooled within-group estimate of the covariance matrix for all categories. Linear discriminant analysis can be further simplified (shrunk) by only using the mean of the diagonal elements of the pooled covariance matrix, i.e., only using the mean of the variable variances and ignoring the between-variable covariances. Classification boundaries based on quadratic discriminant analysis can be curved, whereas classification boundaries based on linear and shrunk-linear discriminant analyses are straight. Crisp classification in linear discriminant analysis can be made on the basis of Mahalanobis distance from the category means, and in shrunk-linear discriminant analysis it can be made on the basis of Euclidian distance. Fig. 2 provides examples of graphical representations of the three different types of covariance matrices, and Fig. 3 provides examples of the classification boundaries based on the three different types of discriminant analysis model (Fig. 3 shows two-dimensional lines, whereas the classification boundaries in the models tested were hyperplanes in a five-dimensional space). The more complex quadratic discriminant analysis will have lower bias than linear discriminant analysis if the covariance matrices of different categories are substantially different, but the latter will usually have lower variance since the pooled covariance estimate is based on more data than the individual-within-category covariance matrices (e.g., if there are three categories then the pooled covariance matrix is estimated on the basis of three times the amount of data). For small to moderate degrees of heteroscedasticity, the smaller variance in the linear model may more than compensate for the larger bias and result in higher correct-classification rates. The shrunk-linear discriminant analysis may have yet greater bias but even less variance.

Regularized discriminant analysis [7, 8] uses a mixture of the estimate of the within-group covariance matrix calculated by pooling data across groups, and the estimates of within-group covariance matrices calculated separately for each group. This allows for the construction of models whose complexity is intermediate between linear and quadratic. Regularized discriminant analysis also allows for additional reduction in model complexity by shrinking the within-group covariance matrices towards the pooled within-group scalar covariance, i.e., the identity matrix

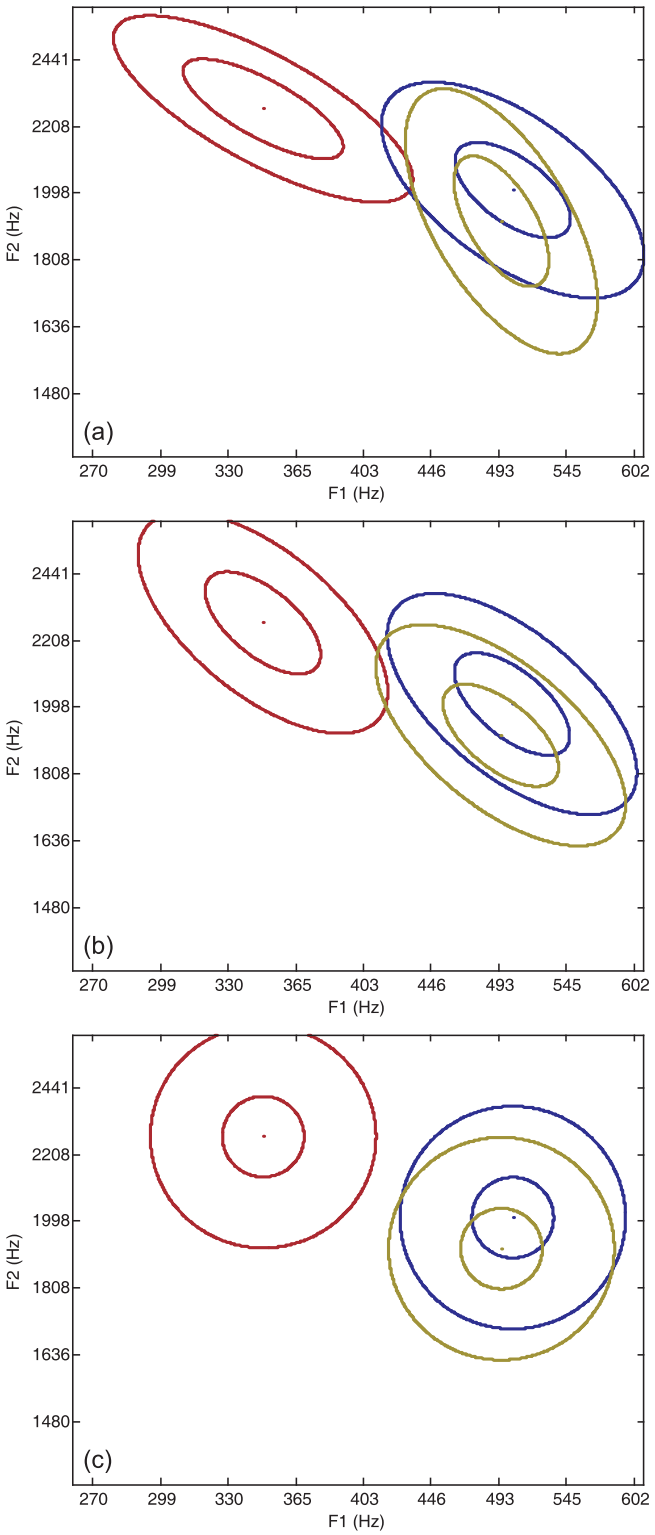

Fig. 2 Graphical representations of (a) examples of the separate covariance matrices for each category used in quadratic discriminant analysis, (b) examples of the pooled covariance matrices used in linear discriminant analysis, and (c) examples of the shrunk covariance matrices used in shrunk-linear discriminant analysis. The red, blue, and yellow ellipsoids represent contours on the probability density functions derived from normalized F1 and F2 at 25% of the duration of tokens of Spanish /i/, /ei/, and /e/ respectively (see Fig. 3).

multiplied by the mean of the diagonal elements of the pooled within-group covariance. The regularized covariance matrices are calculated as in Eq. (1):

$$\boldsymbol{\Sigma}_{REGv} = \alpha\boldsymbol{\Sigma}_v + (1-\alpha)\left(\gamma\boldsymbol{\Sigma} + (1-\gamma)\mathrm{trace}(\boldsymbol{\Sigma})\boldsymbol{I}/p\right) \qquad (1)$$

Where $\boldsymbol{\Sigma}_{REGv}$ is the regularized covariance matrix for vowel category v, $\boldsymbol{\Sigma}_v$ is the covariance matrix estimated using data from category v, $\boldsymbol{\Sigma}$ (with no subscript) is the covariance matrix estimated using data pooled across all vowel categories, $\boldsymbol{I}$ is the identity matrix, p is the number of variables, α is the regularization coefficient (range 0 = linear to 1 = quadratic), and γ is the shrinkage coefficient (range 0 = full shrinkage to 1 = no shrinkage).

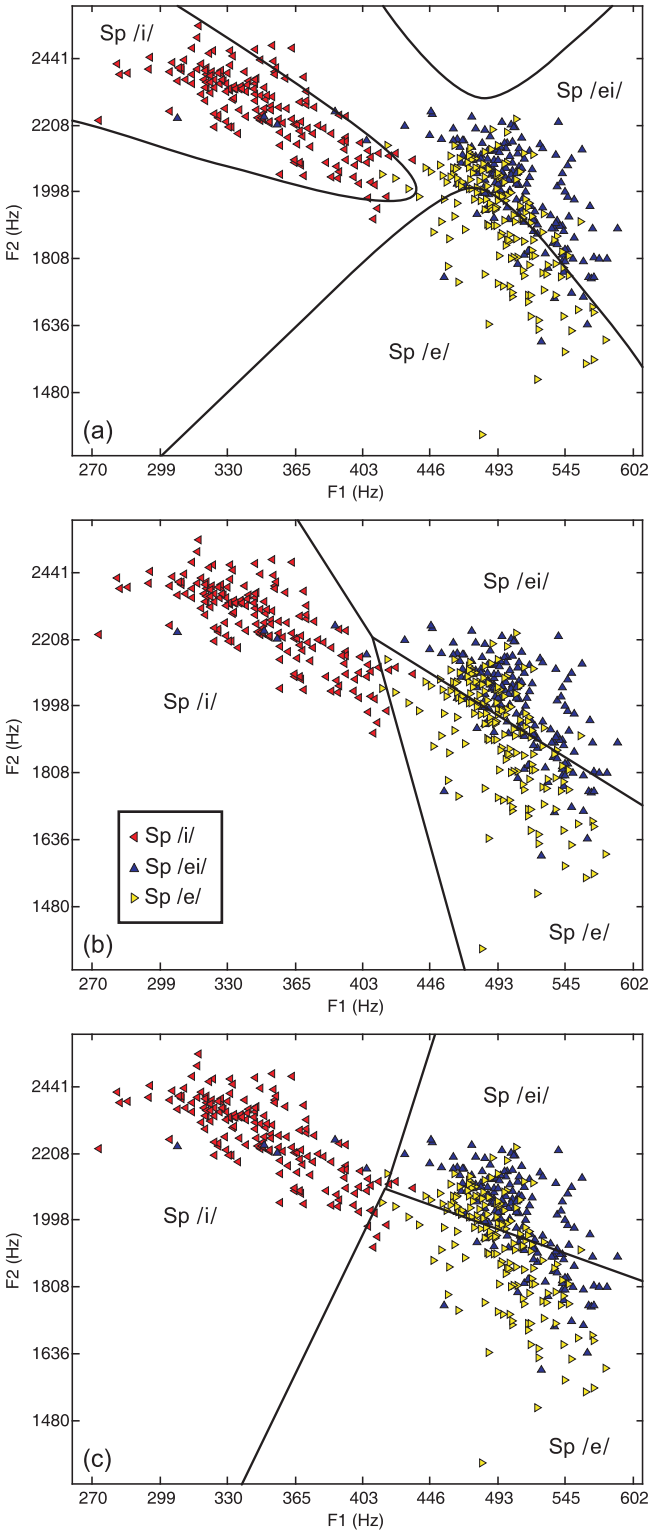

Fig. 3 (a) Examples of curved boundaries based on quadratic discriminant analysis. (b) Examples of straight boundaries based on linear discriminant analysis. (c) Examples of straight boundaries based on shrunk-linear discriminant analysis. The triangles represent the acoustic properties of the vowel tokens used to train the models, and the solid lines represent the categorization boundaries derived by the models. Two predictor variables were included in the models used to create these plots: normalized F1 and F2 at 25% of the duration of the vowel.

4 Results

Regularized discriminant analysis models were fitted to the monolingual Spanish and monolingual English speakers' acoustic data. Appropriate values for the regularization and shrinkage coefficients were determined via leave-one-speaker-out cross-validations. Fig. 4 shows the cross-validated classification-error rates calculated over a fine grid of values for the regularization and shrinkage coefficients. The regularisation and shrinkage coefficient values which resulted in the lowest rates of classification error were $\alpha = 0.12$ and $\gamma = 0.00$ for Spanish, resulting in a correct-classification error of 5.33%, and $\alpha = 0.05$ and $\gamma = 0.02$ for English, resulting in a correct-classification error of only 1.24%.

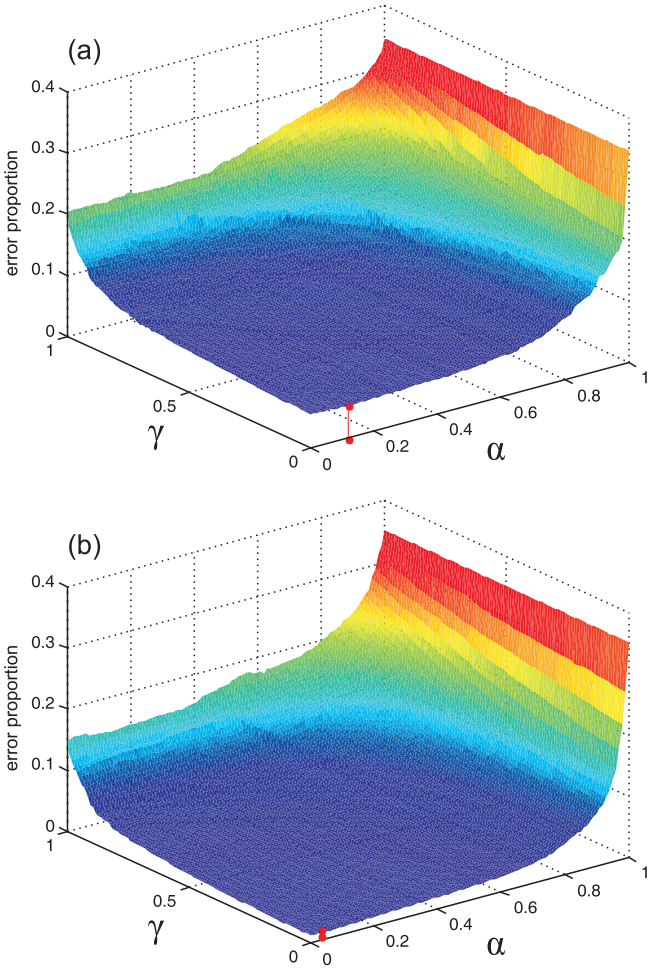

Fig. 4 Cross-validated classification error over a matrix of regularization coefficient values, α, and shrinkage coefficient values, γ. (a) Spanish regularized discriminant analysis model. (b) English regularized discriminant analysis model.

Table 1a provides a confusion matrix of the Spanish regularized discriminant analysis model's cross-validated classification of the Spanish vowel tokens, and Table 1b provides a confusion matrix of the Spanish regularized discriminant analysis model's classification of the English vowel tokens. Most of the first-language classification errors were due to tokens of Spanish /i/ being misclassified as Spanish /e/.

Table 2a provides a confusion matrix of the English regularized discriminant analysis model's cross-validated classification of the English vowel tokens, and Table 2b provides a confusion matrix of the English regularized discriminant analysis model's classification of the Spanish vowel tokens.

5 Discussion and Conclusions

The most noteworthy result of the regularized discriminant analyses is that, for both Spanish and English, the model which resulted in the lowest rate of classification errors was a model which had very small values for the regularization and shrinkage coefficients and was thus close to the least complex model available, i.e., linear discriminant analysis using the mean within-variable variances but not between-variable covariances. This result may be due to the relatively small amount of data available for model training (data from 17 Spanish speakers and 19 English Speakers); however, classification error rates were very low (impressively low in the case of the English model), and it may be that there are constraints on human vowel production and perception which mitigate towards simple patterns for perceptual boundaries [9]. Additional experiments will be necessary to assess the degree of correlation between the statistical models presented here and monolingual Spanish and English listeners' perception of these vowels. If the "simplest works best" finding is upheld, then this opens the possibility of easily obtaining rough-and-ready predictions of cross-language vowel perception on the basis of published summary statistics where only the means and variances of the acoustic properties of categories are provided. Such rough-and-ready predictions could be used as a basis upon which to select subgroups of vowels which may warrant further investigation.

The most noteworthy observations regarding the cross-language predictions made by the regularized discriminant analysis models are that almost all the tokens of English /i/ were classified as Spanish /i/, and the majority of tokens of Spanish /i/ were classified as English /i/, while all the tokens of English /ɪ/ were classified as Spanish /e/, and the majority of tokens of Spanish /e/ were classified as English /ɪ/. English /i/ therefore appears to be similar to Spanish /i/, and English /ɪ/ appears to be similar to Spanish /e/. Spanish learners of English have often been reported to confuse English /i/ and English /ɪ/ [10, 11, 12, 13]; however, the results of the present study suggest that North Central Peninsular Spanish learners of Western Canadian English would initially assimilate tokens of English /i/ to Spanish /i/ and assimilate tokens of English /ɪ/ to Spanish /e/ (similar to Peruvian Spanish listeners' perception of Scottish English vowels [14]), and thus would not be expected to have difficulty distinguishing the two English vowels.

(a)	Classified		
Produced	Sp /i/	Sp /ei/	Sp /e/
Sp /i/	89.4		10.6
Sp /ei/		99.8	1.2
Sp /e/			100

(b)	Classified		
Produced	Sp /i/	Sp /ei/	Sp /e/
Eng /i/	99.5		0.5
Eng /ɪ/			100
Eng /e/		94.3	5.7
Eng /ɛ/			100

Table 1 Confusion matrix of the classification of vowel tokens by the Spanish regularized discriminant analysis model. The values in the cells are the percentage of tokens of the vowel category of the row which are classified as the vowel category of the column. Blank cells have a value of zero. (a) Cross-validated classification of Spanish vowels. (b) Classification of English vowels.

(a)	Classified			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	100			
Eng /ɪ/		99.5		0.5
Eng /e/	0.5		99.5	
Eng /ɛ/		1.0		99.0

(b)	Classified			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Sp /i/	91.8	8.2		
Sp /ei/			100	
Sp /e/	0.6	81.0	7.7	10.7

Table 2 Confusion matrix of the classification of vowel tokens by the English regularized discriminant analysis model. The values in the cells are the percentage of tokens of the vowel category of the row which are classified as the vowel category of the column. Blank cells have a value of zero. (a) Cross-validated classification of English vowels. (b) Classification of Spanish vowels.

Acknowledgments

This research was supported by the Social Sciences and Humanities Research Council of Canada. I wish to thank the speakers who volunteered to take part in the research and all those who assisted with recruitment and facilities in Vitoria-Gasteiz and Edmonton. My thanks to Terry Nearey with whom I discussed many ideas related to this research, and thanks also to Dale Schuurmans who introduced me to regularized discriminant analysis.

Data collection and acoustic analysis took place while the author was at the Department of Linguistics, University of Alberta, the statistical models were built while he was at the Department of Cognitive & Neural Systems, Boston University, and the final writing of the paper took place at the School of Language Studies, Australian National University.

References

[1] P. K. Kuhl, "Cracking the speech code: How infants learn language," *Acoust. Sci. & Tech.* 28, 71–83 (2007)

[2] J. F. Werker, S. Curtin, "PRIMIR: A developmental framework of infant speech processing," *Lang. Learn. & Develop.* 1, 197–234 (2005)

[3] J. M. Hillenbrand, T. M. Nearey, "Identification of resynthesized /hVd/ utterances: Effects of formant contour," *J. Acoust. Soc. Am.* 105, 3509–3523 (1999)

[4] T. M. Nearey, P. F. Assmann, "Modeling the role of vowel inherent spectral change in vowel identification," *J. Acoust. Soc. Am.* 80, 1297–1308 (1986)

[5] S. Zahorian, A. Jagharghi, "Spectral-shape features versus formants as acoustic correlates for vowels," *J. Acoust. Soc. Am.* 94, 1966–1982 (1993)

[6] G. S. Morrison, T. M. Nearey, "A cross-language vowel normalisation procedure", *Can. Acoust.* 34(3), 94–95 (2006)

[7] J. H. Friedman, "Regularized discriminant analysis," *J. Am. Stat. Assoc.* 84, 165–175 (1989)

[8] T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York (2001)

[9] T. M. Nearey, "Speech perception as pattern recognition," *J. Acoust. Soc. Am.* 101, 3241–3254 (1997)

[10] J. E. Flege, O.-S. Bohn, S. Jang, "Effects of experience on non-native speakers' production and perception of English vowels," *J. Phonetics* 25, 437-470 (1997)

[11] P. Escudero, P. Boersma, "Bridging the gap between L2 speech perception research and phonological theory," *Stud. Second Lang. Acq.* 26, 551–585 (2004)

[12] G. S. Morrison, "L1-Spanish speakers' acquisition of the English /i/–/ɪ/ contrast: Duration-based perception is not the initial developmental stage," *Lang. Speech* (In Press)

[13] G. S. Morrison, "L1-Spanish speakers' acquisition of the English /i/–/ɪ/ contrast II: Perception of vowel inherent spectral change," *Lang. Speech* (In Press)

[14] P. Escudero, *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*, PhD diss., University of Utrecht, LOT, Utrecht, The Netherlands (2005)