Morrison, G. S., & Nearey, T. M. (2006). A cross-language vowel normalisation procedure. *Canadian Acoustics, 34(3)*, 94-95.

A Cross-Language Vowel Normalisation Procedure

Geoffrey Stewart Morrison and Terrance M. Nearey
Department of Linguistics, University of Alberta, Edmonton, AB, T6G 2E7

1. Introduction

Vowel classification models trained on production data typically have higher correlation with human listeners' perception when the acoustic properties of the production data are normalised prior to training and testing. Vowel normalisation procedures seek to remove inter-speaker variance due to factors such as vocal tract size, which human listeners discount when identifying vowels. Extrinsic normalisation makes use of information from a representative sample of a speaker's vowel inventory. For example, Nearey's log-mean normalisation [1] applies Equation 1:

$$N_{ktvs} = G_{ktvs} - \overline{G}_s \tag{1}$$

where N_{ktvs} is the normalised value of G_{ktvs} which is the *k*-th formant frequency (in log-Hertz) of instance *t* of vowel category *v* produced by speaker *s*; and $\overline{G}_s$, the within-speaker normalisation factor, is the mean of all the speakers' vowel productions:

$$\overline{G}_s = \frac{1}{(V \cdot T \cdot K)} \cdot \sum_{v=1}^{V} \sum_{t=1}^{T} \sum_{k=1}^{K} G_{ktvs} \tag{2}$$

Normalised vowel formant values are therefore represented as a displacement from a reference value $\overline{G}_s$. This normalisation is valid under the assumptions that the set of vowels used to calculate $\overline{G}_s$ have the same pattern for all speakers, but that vocal-tract size differences may shift the pattern along a track in the F1–F2 space (constant ratio hypothesis / constant log-interval hypothesis). The first assumption can reasonably be expected to be fulfilled when all the speakers share the same language and dialect, but is clearly violated across languages or dialects when they have different numbers of vowels in their inventories, or different skews in the distribution of the same number of vowels, see [2].

This paper presents and tests a variation of log-mean normalisation which may be applied in cross-language and cross-dialect experiments.

2. Normalisation Procedure

Imagine an ideal bilingual who is indistinguishable from a monolingual speaker of language A when speaking language A, and indistinguishable from a monolingual speaker of language B when speaking language B. The ideal bilingual would likely have different $\overline{G}_s$ for language A than for language B, hereafter $\overline{G}_A$ and $\overline{G}_B$. Since vowels from both languages are produced by the same speaker, differences between $\overline{G}_A$ and $\overline{G}_B$ would not be due to differences in vocal tract size. However, $\overline{G}_A$ and $\overline{G}_B$ would differ as a result of cross-language differences in inventory size and skew. Formant values for vowels from language A would be normalised as displacements around $\overline{G}_A$, and vowels from language B as displacements around $\overline{G}_B$. Displacements around the reference value for one language can be translated to displacements around the reference value for the other language by adding or subtracting the difference between $\overline{G}_A$ and $\overline{G}_B$, a cross-language normalisation factor.

Over sufficiently large and balanced samples of speakers from each language, it is reasonable to expect that the mean vocal-tract length (and any other factors underlying formant scale differences) are approximately equal. If so, then rather than relying on mythical ideal bilinguals, the cross-language normalisation factor, $\overline{G}_L$, can be calculated as the difference between the mean reference values from samples of L1 speakers from each language.

$$\overline{G}_L = \overline{\overline{G}}_A - \overline{\overline{G}}_B \tag{3}$$

$$\overline{\overline{G}}_A = \frac{1}{S_A} \cdot \sum_{s_A=1}^{S_A} \overline{G}_{s_A} \qquad \overline{\overline{G}}_B = \frac{1}{S_B} \cdot \sum_{s_B=1}^{S_B} \overline{G}_{s_B} \tag{4}$$

The procedure for classifying vowels from language B in terms of categories from language A is as follows: First, perform a within-language vowel normalisation for vowel formant data from language A using Equation 5.

$$N_{ktvs_A} = G_{ktvs_A} - \overline{G}_{s_A} \tag{5}$$

Second, train a model on the normalised data from language A. Third, apply a cross-language normalisation to vowel formant data from language B using Equation 6.

$$N_{ktvs_A} = G_{ktvs_B} - \overline{G}_{s_B} - \overline{G}_L \tag{6}$$

Finally, use the model trained on language A to classify the cross-language normalised data from language B.

3. TEST OF PROCEDURE

3.1 Method

The effectiveness of the cross-language normalisation procedure was tested using acoustic data from productions of L1-Spanish non-low front vowels (Sp/i/,Sp/ei/,Sp/e/), and L1-English non-low front vowels (Eng/i/,Eng/ɪ/,Eng/e/,Eng/ɛ/). The data, taken from [3], consisted of 10 productions of each vowel category from 59 L1 speakers of various Spanish dialects (32 females and 27 males) and 49 L1 speakers of Western Canadian English (32 females and 17 males).

Linear discriminant analysis models were trained on the L1-English data using the following acoustic variables: F1 and F2 measured at 25% of the duration of the vowel, ΔF1 and ΔF2 (the formant differences between 25 and 75% of the duration of the vowel), and vowel duration. The models were then used to classify the L1-Spanish data. Three models were trained and tested, one used non-normalised log-Hertz values for both training and testing, a second used within-language-normalised values (Equation 1 applied to training and test data), and the third used cross-language-normalised values (Equation 5 applied to training data, Equation 6 applied to test data). Normalisation was applied to both formant and duration measurements. Because of a difference in the ratio of male to female speakers across the language groups, 10 L1-Spanish males were randomly selected, and their data was excluded from the calculation of $\overline{G_L}$; hence, $\overline{G_L}$ was based on a male to female ratio of 17:32 in both languages.

A subset of L1-Spanish vowels (3 instances of each vowel category randomly selected from the productions of each of 28 randomly selected L1-Spanish speakers) were presented to eleven monolingual English listeners for identification (each stimulus was identified once by each listener). The perception experiment also included L1 and L2-English vowels.

The correlations between the listeners' identifications (proportion of responses for each vowel category for each stimulus pooled over listeners, $\boldsymbol{X}$) and the classifications from a model (a posteriori probabilities for each vowel category for each stimulus, $\boldsymbol{Y}$) were calculated using Equation 7, where v indexes the vowel category, V is the number of vowel categories (4), u indexes the stimulus, and U is the number of stimuli (252); see [4, appendix].

$$r = \frac{\left(UV\sum_u\sum_v x_{uv}y_{uv}\right) - \left(\sum_u\sum_v x_{uv}\cdot\sum_u\sum_v y_{uv}\right)}{\sqrt{\left(UV\sum_u\sum_v x_{uv}^2 - \left(\sum_u\sum_v x_{uv}\right)^2\right)\cdot\left(UV\sum_u\sum_v y_{uv}^2 - \left(\sum_u\sum_v y_{uv}\right)^2\right)}} \quad (7)$$

3.1 Results

The models trained on normalised L1-English vowels had a slightly higher correct-classification rate on the training data than the model trained on non-normalised vowels: 98.7% versus 96.2% in a leave-one-participant-out cross-validation.

For the test data, correlation with the listeners' perception of L1-Spanish vowels was greater for the cross-language-normalised model, r = .869, than for the within-language-normalised, r = .853, and the non-normalised models, r = .848. Pooled confusion matrices are given in Tables 1 and 2.

Table 1. L1-English listeners' identification of L1-Spanish vowels, expressed as proportions pooled across repetitions, speakers, and listeners.

Produced	Perceived			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Sp /i/	.951	.036	.009	.004
Sp /ei/	.005	.003	.982	.010
Sp /e/	.004	.275	.473	.248

Table 2. Mean a posteriori probabilities for classification of L1-Spanish vowels by the cross-language normalised model.

Produced	Classified			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Sp /i/	.997	.001	.001	
Sp /ei/			1.000	
Sp /e/	.014	.583	.286	.117

4. CONCLUSION

Use of the cross-language vowel normalisation procedure increased the correlation between monolingual English listeners' perception of L1-Spanish vowels and the classification of L1-Spanish vowels by a statistical model trained on L1-English vowel productions.

REFERENCES

[1] Nearey, T. M. & Assmann, P. F. (in press). "Probabilistic 'sliding-template' models for indirect vowel normalization," in *Experimental Approaches to Phonology*, edited by M. J. Solé, P. S. Beddor, and M. Ohala (Oxford: Oxford University Press).

[2] Disner, S. F. (1980). "Evaluation of vowel normalization procedures," *J. Acoust. Soc. of Am., 67*, 253–261.

[3] Morrison, G. S. (2006). "L1 & L2 production and perception of English and Spanish vowels: A statistical modelling approach," unpublished doctoral dissertation, University of Alberta.

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

This work was supported by SSHRC.

This page is an addendum to the version of this paper published as:
Morrison, G. S., & Nearey, T. M. (2006). A cross-language vowel normalisation procedure. *Canadian Acoustics, 34(3)*, 94–95.

ABSTRACT

Extrinsic vowel normalisation procedures, such as log-mean normalisation, are appropriate when all speakers share the same language and dialect. In such a situation, it is reasonable to assume that all speakers will have the same pattern of distribution of vowels: Each speaker has the same patten of displacements from their own reference value. The reference value for each speaker is based on the mean formant values of all of that speaker's vowels. The normalisation procedure shifts the pattern along a track in the F1-F2 space, with the aim of removing variance due to factors such as vocal tract length (constant ratio hypothesis / constant log-interval hypothesis). However, in cross-language or cross-dialect situations, there may be differences in vowel inventory size or differences in skew of the distributions of the same number of vowels, which can lead to differences in the reference values. This paper presents a variant of log-mean normalisation which can be applied in cross-language or cross-dialect experiments. The effectiveness of the cross-language normalisation procedure was tested: Three statistical models were trained on L1-English vowel data and then used to classify L1-Spanish vowels. One model was trained and tested on non-normalised data, a second was trained and tested on within-language normalised data, and the third was tested on cross-language normalised data. The correlation between monolingual English listeners' identification of Spanish vowels, and the models' classifications of the same Spanish vowels was greatest for the cross-language normalised model.

Post-Publication Clarification:

For calculating a speaker's $\overline{G}_s$ for a given language ($\overline{G}_A$ or $\overline{G}_B$), each vowel category within the language was given equal weight. This is a point worth making since in some cases we did not have exactly 10 samples of each vowel from each speaker: For example, if we had9 /i/ tokens from one speaker and 11 from another, we calculated the mean /i/ value for each speaker from all the data available from that speaker then calculated the $\overline{G}_s$ for all the speaker's vowels. Also some vowels are longer than others, and a vowel token's mean value was calculated over the vowel formant track from 25–75% of the duration of the vowel, so the mean for each vowel token was calculated before calculating the mean for the corresponding vowel category for that speaker. In evaluating the procedure, we therefore applied a revised version of Equation 2:

(2 revised)

$$\overline{G}_s = \frac{1}{V}\sum_{v=1}^{V}\left(\frac{1}{T_{vs}}\sum_{t=1}^{T_{vs}}\left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{R_{tvs}}\sum_{r=1}^{R_{tvs}} G_{rktvs}\right)\right)\right)$$

Where:

r indexes a point along a formant track

R_{tvs} is the number of points on the formant tracks of token t of vowel category v of speaker s (the same number of points are available for each formant k)

T_{vs} is the number of tokens of vowel category v of speaker s

All other symbols are as in the paper. The same number of formants, K, are measured for each vowel token. The same number of vowels, V, are produced by each speaker.

Alternative version:

An alternative to calculating $\overline{G}_L$ is to add back in the mean reference value for the speaker's L1:

$$N_{ktvs} = G_{ktvs_A} - \overline{G}_{s_A} + \overline{\overline{G}}_A \qquad \text{(5 alt.)}$$

$$N_{ktvs} = G_{ktvs_B} - \overline{G}_{s_B} + \overline{\overline{G}}_B \qquad \text{(6 alt.)}$$