A comparison between human vowel normalization strategies and acoustic vowel transformation techniques

Patti Adank, Roeland van Hout, and Roel Smits*

Department of Linguistics
University of Nijmegen, the Netherlands
* Max Planck Institute for Linguistics, University of Nijmegen, The Netherlands
p.adank@let.kun.nl

Abstract
Perceptual and acoustic representations of vowel data were compared directly to evaluate the perceptual relevance of several speaker normalization transformations. The acoustic representations consisted of raw F0 and formant data. The perceptual representations were obtained through an experimental procedure, with phonetically trained listeners as subjects. The raw acoustic data were transformed according to several normalization schemes. The perceptual and the acoustic representations were compared using regression techniques. A z-score-transformation of the raw data appeared to resemble the perceptual data.

1. Introduction
Speaker normalization can be regarded as the process by which listeners perceptually remove variation due to speaker differences when listening to spoken language.

Since the early 1950s, phoneticians have tried to understand the normalization process by attempting to mimic it in the acoustic domain by, for instance, warping the acoustic representations of vowel stimuli (F0, formants) into different perceptual scales (e.g. bark [1]) or transforming the F0 and formant values to z-scores [2]. For detailed reviews of the speaker normalization issue and normalization transformations see [3].

Several studies have made comparisons between the proposed normalization methods in order to establish their relative perceptual relevance. In general, three approaches have been taken in making such comparisons. The first involves comparing transformed acoustic measurements directly to other (transformed) measurements [4]. A normalization procedure is judged on how well it reduces speaker-specific variance in the data. A second approach entails comparing the transformed acoustic data with perceptual judgments of phonetically trained listeners on the language-specific characteristics of the vowel data in question [5] to see how well the clustering of transformed data agrees with these (indirect) judgments. A third type evaluates the perceptual relevance of normalization schemes by classifying them on the basis how well their output data can be categorized into discrete vowel categories correctly [1, 6]. The third kind seems to approach human speaker normalization closest in that it attempts to use some kind of pattern recognition. It involves mapping continuous acoustic data onto discrete perceptual units (vowel categories). However, it is likely that the perceptual speaker normalization process involves more than labeling the vocalic input into discrete vowel categories alone, since differences in linguistic quality (i.e. more open articulated) between spoken vowels of the same category can be perceived by humans too. Human perception is therefore more continuous in nature than can be accounted for by the use of a discrete perceptual scale. Also, perceptual studies have showed that only a small percentage of the vowels are misclassified or confused in vowel labeling tasks [7]. The resulting classifications are therefore not suitable for detailed perceptual modeling.

Comparing a continuous acoustic data with continuous perceptual data could possibly account for the relevant differences within vowel categories and could therefore model normalization better than using a discrete perceptual scale.

The aim of this study is to model speaker normalization by listeners. This will be attempted by comparing continuous acoustic data with continuous perceptual data. The perceptual data consisted of perceptual correlates of articulatory dimensions: judgements of listeners of each vowel’s height, tongue advancement. The acoustic data consisted of the traditional acoustic correlates of articulatory dimensions: measurements of F0 and F1, F2 and F3. The acoustic data were transformed following several normalization schemes and compared to the perceptual data directly using discriminant analysis and regression techniques.

The perceptual data were obtained through an experimental procedure in which phonetically trained listeners were asked to assign a label to a vowel stimulus and to judge its linguistic quality on three dimensions on a continuous scale.

2. Perception experiment
2.1. Speech material
The speech material consisted of recordings of a socio-economically homogeneous group of 20 speakers of Dutch (i.e. same education level, originating from and working and living in the same region in the Netherlands), who varied as much as possible anatomically-physiologically. The speaker group consisted of five younger females (22-44 years old), five older females (45-60 years old), five younger males and five older males. For each of these speakers recordings of all nine Dutch monophthong vowels: /aI/, /a/, /E/, /I/, /U/, /i/, /o/, /U/, and /y/ in a neutral /sVs/ context were available. In total 180 /sVs/ stimuli were recorded. In total 180 /sVs/ stimuli were recorded.

The recordings were made on a TASCAM DA-P1 portable DAT-recorder, with an AKG C420 Headset condenser microphone. The syllables used in this experiment were downsampled to 16 kHz.

2.2. Subjects
Twelve phonetically trained (expert) listeners participated in the experiment. One of them participated only in a pilot version of the experiment, while the other 11 participated in
the actual experiment. The subjects who were asked to participate in this experiment were selected on the basis of their known extensive experience with narrow transcription of speech sounds. Also, all of them had received formal training in phonetic transcription using the IPA system.

2.3. Experimental interface

Figure 1 shows the interface used in the experiment. It contains:

- Nine vowel buttons for phonetic labeling, corresponding to the nine monophthong vowels of Dutch. If the subject clicked on one of them, the corresponding vowel category was recorded.
- A copy to scale of the IPA vowel quadrilateral for judging vowel height and backness. If the subject clicked her/his mouse inside the quadrilateral, the coordinates corresponding to that place were written to the response file. The horizontal axis of the quadrilateral represented tongue advancement (from left to right representing front to back). The vertical axis represented vowel height (from bottom to top representing low to high perceived vowel height).
- A rectangular field for judging lip rounding/spreading. The plus sign indicates a maximally rounded vowel; the minus sign stands for maximally spread. If the subject clicked the mouse on the axis, the coordinate corresponding to the horizontal axis was recorded.

![Figure 1: Experimental interface used in the experiment.](image)

The subjects were told that the corners of the quadrilateral as well as the end points of the rounding/spreading axis were to be regarded as the theoretical end points for vowel height, tongue advancement and lip rounding/spreading respectively.

2.4. Experimental procedure

The procedure of one experimental cycle (in which one vowel token in a /sV/-context was judged) was as follows. First, a signal tone was played (a sine wave of 400 Hz and with a duration of 100 ms). Next, the stimulus was presented up to ten times with 1.5-second intervals. While the stimulus was repeated, the subject had to do three things: first she/he had to label the vowel by identifying the vowel by clicking on one of the vowel buttons (1), second he/she had to judge the rounding/spreading of the vowel by clicking on the axis in the rounding/spreading rectangle (3) and third the subject had to judge the vowel’s height and tongue advancement by clicking in the vowel quadrilateral. After the subject had performed the three tasks, the experimental program would proceed to the next experimental cycle, regardless how often the stimulus had been presented. Each experiment was preceded by nine familiarization stimuli from a speaker who did not belong to the 20 speakers whose vowels were to be judged.

For three speakers the vowels were judged twice, in order to be able to assess the reliability and consistency of each subject. In all, 207 judgments (20 * three speakers * nine vowel categories) were obtained in the experiment. Stimuli were presented blocked per speaker.

2.5. Results

Cohen's kappa was calculated for the 27 vowels that were judged twice in the experiment in order to establish the intrarater consistency in labeling. A value of 1 indicates that the 27 vowels were assigned to the same vowel category on the two occasions these stimuli had to be judged. Eight out of 11 subjects had values of Cohen's kappa >0.9, thus indicating high intrarater consistency. The three other subjects (1, 7 and 9) showed values between 0.8 and 0.9, indicating slightly lower, but still high, consistency.

The second step was the inspection of the mismatches between the stimulus vowel category (as intended by the speaker) and the chosen vowel category (labeled by the expert listeners). The overall number of mismatches was surprisingly low (5.3%). However, two out of the 11 subjects (again 1 and 7) showed a considerable higher number of mismatches than the other nine subjects. After further inspection of the mismatches, it was decided to exclude subjects 1 and 7 from further analysis, since their judgments appeared to be the result of different strategies than the other nine subjects.

Only the results for the 180 vowels that were judged once were selected for further analysis.

The mean results per stimulus vowel for Height * Advancement (nine judgments per stimulus vowel) are shown in figure 2.

![Figure 2: Height * Advancement judgments for the 180 stimulus vowels pooled for the nine subjects.](image)
specific variation can be seen. There appear to be some differences in vowel quality within vowel categories.

3. Acoustic data

The speech material was the same as described in section 2.1. \( F_1 \) through \( F_3 \) were measured by hand for all 180 stimulus vowels using SpeechStation II, a program for sound spectrography. While measuring, the number of LPC-coefficients (used for the fitting of a spectral envelope) was adjusted for each speaker, as well as the ceiling frequency. The values were measured approximately at the vowel’s temporal center point.

\( F_0 \) was estimated automatically, using an autocorrelation method at the vowel’s temporal center point. The allowed range was 50-300 Hz for male speakers.

In figure 3, a scatterplot of the first and second formant frequencies of the 180 stimulus vowels is displayed in hertz.

The raw data show patterns typically found in \( F_1-F_3 \) plots [8]. The vowel categories show wide and overlapping clusters, unlike the distribution in figure 1, which shows strong clustering per vowel category.

The raw \( F_0 \) and formant data were transformed using several transformation schemes. These are: scale (Hz, log, bark and ERB) as independent (predicting) variables. The mean values per stimulus vowel were used as the perceptual (grouping) variable. The mean values per stimulus vowel (pooled for the nine subjects) for Height, Advancement and Rounding were used as the perceptual values. The results are displayed in table 2.

Table 2 shows that the perceptual data can be grouped into vowel categories without a single confusion. The experts appear to have used a very distinct separation between vowel categories. The acoustic data show more overall confusion than the perceptual data. Still, categorization is relatively high, especially for the raw data (\( F_0, F_1-F_3 \) in hertz). The z-transformed data appears to perform best, followed by the logmean transformation. Both normalisations appear to perform considerably better than the different scales (bark, raw log(Hz) and ERB), that show almost no improvement relative to the raw, untransformed, data.

Table 3: Values for \( R^2 \) for the linear regression analyses for the three criterion variables with the scale-transformed data as predictor variables

<table>
<thead>
<tr>
<th>criterion</th>
<th>Hertz</th>
<th>Log (hertz)</th>
<th>Bark</th>
<th>ERB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>0.730</td>
<td>0.719</td>
<td>0.738</td>
<td>0.734</td>
</tr>
<tr>
<td>Advancement</td>
<td>0.875</td>
<td>0.900</td>
<td>0.900</td>
<td>0.900</td>
</tr>
<tr>
<td>Rounding</td>
<td>0.614</td>
<td>0.577</td>
<td>0.589</td>
<td>0.586</td>
</tr>
</tbody>
</table>

The portion of explained variance seems to be highest for Advancement; it is slightly lower for Height and Rounding. The four scales do not show much difference. For Height, only two out of three show higher \( R^2 \)’s than the raw, baseline, data. For Advancement, all three show higher values than the raw data and for rounding all three are lower. Overall, no clear pattern can be seen in the relative performance for the four scales.

The regression equations for the three criterion variables for the raw data in hertz are (only predictors with a significance level of p<0.01 are included):

\[ \text{Height} = -0.3F_0 + 0.78F_1 - 0.2F_2 + 14.34 \]
\[ \text{Advancement} = 0.14F_1 + 0.92F_2 + 0.19F_3 + 80.72 \]
\[ \text{Rounding} = 0.33F_0 - 0.33F_1 - 0.72F_2 + 134.7 \]

When looking at the \( \beta \)-coefficients for Height, it can be seem that \( F_1 \) in hertz, the correlate that is traditionally thought to correlate with vowel height, is the highest. The same pattern can be seen in the coefficients of Advancement; here \( F_2 \) in hertz is the highest. For rounding, the pattern is less clear. It could be expected that only \( F_3 \) should be significant, since \( F_3 \) is traditionally associated with rounding, but instead \( F_3 \) is not significant.

![Figure 3: F2 * F1 for the 180 stimulus vowels in hertz.](image)
The second series of regression analyses was performed on the acoustic data that were transformed according to the zscore and logmean normalization schemes. The results are shown in table 4.

Table 4: Values for R² for the linear regression analyses with the zscore-transformed data and the logmean data.

<table>
<thead>
<tr>
<th>criterion</th>
<th>F₀, F₁, F₂, F₃ in zscores</th>
<th>F₀m, F₁m, F₂m, F₃m in Log(hertz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>0.801</td>
<td>0.735</td>
</tr>
<tr>
<td>Advancement</td>
<td>0.914</td>
<td>0.915</td>
</tr>
<tr>
<td>Rounding</td>
<td>0.732</td>
<td>0.669</td>
</tr>
</tbody>
</table>

The same patterns can be observed in table 4 and 3 in the values of R² with respect to the three criterion variables. The zscore transformation shows higher than the logmean values. When the values for the zscores are compared with the values for the three scale transformation in table 3, the zscore appears to show higher values than the three scale transformation for all three criterion variables.

The regression equations for the three criterion variables for the data in zscores in hertz are (only predictors with a significance level of p≤0.01 are included):

Height  = -0.15zF₀ + 0.75zF₁ - 0.18zF₂ + 41.56
Advancement = 0.14zF₁ - 0.92zF₂ + 0.17zF₃ + 50.39
Rounding  = -0.35zF₁ - 0.73zF₂ - 0.27zF₃ + 46.62

The β-coefficients for Height, show that, again, F₁ in hertz is the highest. For Advancement F₂ is the highest. For rounding, the pattern still unclear, although the equation shows a significant F₃.

The regression equations for the three criterion variables for the log-transformed data are (only predictors with a significance level of p≤0.01 are included):

Height  = -0.35logF₁ - 0.158logF₂ - 0.37logF₃ - 34.22
Advancement = 0.27logF₀ - 0.92logF₁ + 0.14logF₂ + 134.61
Rounding  = -0.32logF₁ - 0.70logF₂ - 0.46logF₃ + 75.61

Again, the same patterns can be observed in the value of the β-coefficients as in the equations for hertz and the zscores, but this time the mean values per speaker in log(Hz) are significant as well, in accordance with the traditional notions about acoustic correlates of vowel quality. For Height, the logmean of F₁ is significant, for Advancement the logmean of F₂ is significant and for rounding the logmean of the F₃.

5. Discussion and conclusion

The results of the linear discriminant and regression analyses show that the normalization scheme using a transformation of all values of the vowels of a speaker to zscores is the most successful. This is true for the pattern classification test (the discriminant analyses, table 2) and for the direct comparison between the perceptual and acoustic representations (regression analysis, table 4). The zscore transformation shows better results than the baseline data (F₀ and formants in hertz), the three scale transformations and the logmean transformation.

The zscore transformation’s success probably can be explained in that it incorporates a correction for the differences between the exact location and shape of each speaker’s complete vowel space. The logmean transformation shows a similar speaker-specific approach: it uses a speaker-specific reference point for the speaker’s vowel space. The results show that the transformation of the entire vowel space is more successful than using a single reference point per speaker.

The zscore and the logmean schemes both perform better than the three scale transformations. This can probably be explained by the fact that the scale transformations incorporate no information about the speaker.

With regard to the different perceptual scales must be concluded that modeling works better for perceived vowel height and tongue advancement than for rounding. More study is necessary to model vowel rounding better.

In conclusion, it has proven to be possible to model speaker normalization by listeners by comparing continuous perceptual data with acoustic representations. Also, comparing continuous perceptual scales with continuous acoustic scales, in addition to a comparison between a discrete perceptual scale and continuous acoustic scales, has proven very useful for the determination of the perceptual relevance plus the variance-reducing capacities of different normalization methods.

6. References