Personalized Quantification of Voice Attractiveness in Multidimensional Merit Space

Yasunari Obuchi
School of Media Science, Tokyo University of Technology
obuchyiysnr@stf.teu.ac.jp

Abstract
Voice attractiveness is an indicator which is somehow objective and somewhat subjective. It would be helpful to assume that each voice has its own attractiveness. However, the paired comparison results of human listeners sometimes include inconsistency. In this paper, we propose a multidimensional mapping scheme of voice attractiveness, which explains the existence of objective merit values of voices and subjective preference of listeners. Paired comparison is modeled in a probabilistic framework, and the optimal mapping is obtained from the paired comparison results on the maximum likelihood criterion.

The merit values can be estimated from the acoustic feature using the machine learning framework. We show how the estimation process works using real database consisting of common Japanese greeting utterances. Experiments using 1- and 2-dimensional merit spaces confirm that the comparison result prediction from the acoustic feature becomes more accurate in the 2-dimensional case.

Index Terms: voice attractiveness, paired comparison, preference, multidimensional, acoustic feature

1. Introduction
Can artificial intelligence (AI) understand human emotions? The growth of the AI industry has raised that question among ordinary people, and the answer would be, “Yes, if it was taught to”.

In the field of speech science, many studies have been done to realize the system that recognizes how the speaking person feels from the voice signal [1, 2]. In these studies, the existence of the training corpus was essential. In fact, that is the way to teach AI what the emotion is.

When creating the corpus, there are two ways of attaching emotion labels to each voice signal. One is to ask a voice actor to pretend as if speaking with a specific emotion, and record the voices. The problem of this method is that sometimes the emotion is too exaggerated. The other way is to ask several listeners to pretend as if speaking with a specific emotion, and record the speakers. Mean opinion score (MOS) test is the most popular method [4]. However, the paired comparison test is also used [5, 6] to lower the listener’s burden and obtain more reliable results.

Even with paired comparison, our final target is the absolute value for each voice. If all comparison results are consistent, we can at least obtain the order of the voices. However, the results include inconsistency in many cases, and a probabilistic interpretation is required. There have been many studies applying probabilistic interpretation of paired comparisons to real world problems, such as in sports events [7].

A typical approach of probabilistic modeling is based on the log likelihood function,

\[
L = \sum_i \sum_j w_{ij} \log f(d_{ij})
\]

where \(w_{ij}\) is the number of times the voice \(i\) was preferred to voice \(j\). In the Bradley-Terry model [8], the probability is defined as though two voices are competing for the shared resource as follows.

\[
f(d_{ij}) = \frac{e^{a_i}}{e^{a_i} + e^{a_j}} = \frac{1}{1 + e^{-d_{ij}}}
\]

In the Thurstone-Mosteller Case V model [9], each voice is assumed to have Gaussian observation probability whose mean corresponds to its own attractiveness. In this case,

\[
f(d_{ij}) = \frac{1}{2} (1 + \text{erf}(d_{ij}))
\]

In both cases, the scaling factor for \(d_{ij}\) was omitted because the attractiveness itself has the freedom of scaling.

After defining the log likelihood, \(\{a_i\}\) can be estimated by maximizing \(L\).
2.2. Personalized Attractiveness Model

In the paired comparison test, two listeners may have opposing opinions for any given pair. Many conventional models interpret it simply as an occurrence of a less likely event. However, it can be interpreted that those two listeners have different criteria regarding voice attractiveness. The variety of criteria can be visualized by using a multidimensional merit space, in which the items are given as points and the preferences are given as direction or weight vectors. There have been some studies proposing a multidimensional extension of the Bradley-Terry model. Fujimoto et al. [10] proposed a mixture model and applied it as a visualization method to the movie rating task. Causeur and Husson [11] proposed a two-dimensional model representing ranking and relevance axes and applied it to the consumer’s preference for cornflakes. Our approach could be regarded as an extension of Fujimoto’s model, in which each listener has a preference direction and the voice attractiveness is defined as the inner product of the merit and preference.

In the proposed model, we introduce the listener dependency into eq. (1), and obtain,

$$L = \sum_{i} \sum_{j} \sum_{k} w_{ijk} \log f(d_{ijk})$$

(4)

where $w_{ijk}$ is the number of times the listener $k$ prefers the voice $i$ to the voice $j$ (either 0 or 1 in most cases). The definition of $d_{ijk} = a_{ik} - a_{jik}$ does not change, where the attractiveness for the listener $k$ is defined as,

$$a_{ik} = p_k \cdot m_i$$

(5)

where $p_k$ ($|p_k| = 1$) is the preference vector for the listener $k$ and $m_i$ is the merit vector for the voice $i$.

Maximization of eq. (4) can be done using the gradient ascent method. In the two-dimensional case, $p_k = (\cos \theta_k, \sin \theta_k)^T$ where $\theta_k$ is called the preference angle, and $m_i = (\xi_i, \eta_i)^T$ where $\xi_i$ and $\eta_i$ are called merit values. The gradient can then be obtained by differentiating $L$ in terms of these three variables. (Note that $f(d_{ijk}) + f(d_{jik}) = 1$)

$$\frac{\partial L}{\partial \xi_i} = \sum_{j \neq i} \sum_k f'(d_{ijk})(\frac{w_{ijk}}{f(d_{ijk})} - \frac{w_{jik}}{1 - f(d_{jik})}) \cos \theta_k$$

(6)

$$\frac{\partial L}{\partial \eta_i} = \sum_{j \neq i} \sum_k f'(d_{ijk})(\frac{w_{ijk}}{f(d_{jik})} - \frac{w_{jik}}{1 - f(d_{jik})}) \sin \theta_k$$

(7)

$$\frac{\partial L}{\partial \theta_k} = \sum_i \sum_j f'(d_{ijk})(\frac{w_{ijk}}{f(d_{jik})} - \frac{w_{jik}}{1 - f(d_{jik})}) r_{ijk}$$

(8)

$$d_{ijk} = (\xi_i - \xi_j) \cos \theta_k + (\eta_i - \eta_j) \sin \theta_k$$

(9)

$$r_{ijk} = (\eta_i - \eta_j) \cos \theta_k - (\xi_i - \xi_j) \sin \theta_k$$

(10)

If we have a limited number of comparison results, it is reasonable to put constraints on $\xi_i$, $\eta_i$, and $\theta$ to remain in a given range. For simplicity, we assume that the merit values are constrained in $[0, 1]$. We also assume that $\xi_i$ and $\eta_i$ have nonnegative influence for the attractiveness for all listeners. Accordingly, the range of $\theta$ is limited in $[0, \pi/2]$. Taking these constraints into account, the maximization algorithm is described in Fig. 1.

As for $f$, we use the Thurstone-Mosteller Case V model (3), and its derivative is given by

$$f'(d_{ijk}) = \frac{1}{\sqrt{\pi}} e^{-a^2_{ijk}/2}$$

(11)

1. repeat
2. initialize $\{\xi_i\}, \{\eta_i\}, \{\theta_k\}$ randomly
3. repeat
4. for all $i, k$
5. calculate $\frac{\partial L}{\partial \xi_i}, \frac{\partial L}{\partial \eta_i}, \frac{\partial L}{\partial \theta_k}$ using (6),(7),(8)
6. end for
7. for all $i, k$
8. $\xi_i \leftarrow \xi_i + \frac{\partial L}{\partial \xi_i}$
9. $\eta_i \leftarrow \eta_i + \frac{\partial L}{\partial \eta_i}$
10. $\theta_k \leftarrow \theta_k + \frac{\partial L}{\partial \theta_k}$
11. end for
12. until converge
13. calculate $L$ using (4) and store $\{\xi_i\}, \{\eta_i\}, \{\theta_k\}, L$
14. until $N$ times
15. return $\{\xi_i\}, \{\eta_i\}, \{\theta_k\}$ that yielded the largest $L$

Figure 1: Pseudocode for log likelihood maximization.

<table>
<thead>
<tr>
<th>Table 1: List of low level descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy/Pitch</td>
</tr>
<tr>
<td>RMS energy</td>
</tr>
<tr>
<td>Log energy</td>
</tr>
<tr>
<td>F0</td>
</tr>
<tr>
<td>Log centroid</td>
</tr>
<tr>
<td>Voicing prob.</td>
</tr>
<tr>
<td>Log entropy</td>
</tr>
<tr>
<td>Harmonicity</td>
</tr>
<tr>
<td>Variance</td>
</tr>
<tr>
<td>Sharpness</td>
</tr>
</tbody>
</table>

3. Merit Estimation from Acoustic Features

Once we have a set of voices with the correct merit value labels, the relationship between the merit values and acoustic features can be extracted using the machine learning framework.

In this paper, we prepare a redundant set of acoustic features, and try to find the optimal subset to estimate the merit values experimentally.

Acoustic features are extracted using OpenSMILE [12] and Julius [13]. OpenSMILE works in a two-stage process. In the first stage, voiceic input is divided into 25 ms overlapping frames with 10 ms frame shift, and various features (called low level descriptors: LLDs) are extracted. In the second stage, LLDs of the same kind are collected from all frames, and various features (called functionals) are extracted. Tables 1 and 2 show the categorical list of LLDs and functionals. A combination of 14 LLDs and 23 functionals resulted in 322 features. More details on OpenSMILE can be found in the online manual.

Julius is an open-source speech recognition software, and it can be used to analyze a voice signal using the forced-alignment function. We calculate the acoustic model score, total utterance length, and the length of the final phoneme (mostly vowels in Japanese). Therefore, our baseline feature set consists of 325 features.

Machine learning is conducted using WEKA [14]. Since we want to predict real values, SMOreg [15], which is one of the state-of-the-art SVM-based regression algorithms was chosen.

Backward stepwise selection (BSS) is used to reduce the number of features effectively\(^1\). From the baseline feature set, one feature is removed after the exhaustive test of every possible

\(^1\)Our preliminary experiments showed that BSS is much better than forward stepwise selection (FSS).
removal. The same procedure is repeated until only one feature is left, and the subset that gets the highest score becomes the optimal subset.

Finally, the model trained by the optimal subset is used to predict the merit values of unknown voice signals. If the listener’s preference angle is known, the attractiveness for the listener can be estimated, and the comparison result can be predicted.

4. Experimental Results

4.1. Recordings and Comparisons

The proposed method was evaluated using a real database. We recorded voices of the Japanese greeting “Irasshaimase (welcome)” uttered by 115 students (44.1kHz sampling, stereo) using a Panasonic RR-XS355 digital voice recorder. “Irasshaimase” is the phrase given by the shop clerk every time a customer enters, so its impression is very important for the business. OpenSMILE used the data in its original format, and Julius used the version converted to 16kHz/monaural.

Eighteen listeners participated in the comparison experiments, in which the listener chose the most attractive greeting among three randomly selected ones after listening to them. Triplet comparison was used instead of paired comparison, to obtain more comparison results with a smaller number of trials. If voice \( A \) was chosen from \( \{ A, B, C \} \), it was interpreted that \( A \) won in the comparisons \( \{ A, B \} \) and \( \{ A, C \} \). Each listener completed 38 or 39 triplet comparisons, and finally, 1,388 paired comparison results were obtained.

4.2. Mapping to a 1- or 2-Dimensional Merit Space

After converting 1,388 comparison results to \( w_{ijk} \), we can maximize \( L \) of (4) in terms of \( p_k \) and \( m_i \). Figure 2 shows the optimal mapping obtained with \( N = 80 \) and \( s = 0.05 \). One-dimensional optimization was also executed with \( N = 100 \) and \( s = 0.002 \) for comparison, in which \( \{ w_{ijk} \} \) was marginalized to \( \{ w_{ij} = \sum_k w_{ijk} \} \) and optimization was done with \( a_c \). The results are shown in Figure 3.

The efficiency of mapping was evaluated using Kendall’s tau-b:

\[
\tau_b = \frac{N_C - N_D}{\sqrt{N_C + N_D + N_T} \sqrt{N_C + N_D}}.
\]

\( N_C \) (\( N_D \)) is the number of concordant (discordant) pairs, in which the voice with the larger (smaller) attractiveness was preferred. \( N_T \) is the number of tied pairs, in which two voices have the same attractiveness. Since our comparison was not exhaus-

![Figure 2: Optimal mapping of 2-dimensional merit space. The voices are represented by circles. There are two listeners with \( \theta_k = 0 \), represented by the x-axis. There are three listeners with \( \theta_k = \pi/2 \), represented by the y-axis. Other listeners are represented by the dashed lines, each of which has the angle \( \theta_k \) to the x-axis.](image)

![Figure 3: Optimal mapping of 1-dimensional merit space. The variable \( a \) is the 1-dimensional merit value. There are 26 voices mapped to \( a = 0 \), and 22 voices mapped to \( a = 1 \).](image)

Table 2: List of functionals. Linreg stands for linear regression, qreg for quadratic regression, and seglen for segment length.

<table>
<thead>
<tr>
<th>Extremes</th>
<th>Regression</th>
<th>Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>max</td>
<td>linreg slope</td>
<td>number of seg</td>
</tr>
<tr>
<td>min</td>
<td>linreg offset</td>
<td>seglen mean</td>
</tr>
<tr>
<td>range</td>
<td>linreg linear error</td>
<td>seglen max</td>
</tr>
<tr>
<td>max position</td>
<td>linreg quadratic error</td>
<td>seglen min</td>
</tr>
<tr>
<td>min position</td>
<td>qreg coef 1</td>
<td>seglen std. dev.</td>
</tr>
<tr>
<td>mean</td>
<td>qreg coef 2</td>
<td></td>
</tr>
<tr>
<td>max–mean</td>
<td>qreg coef 3</td>
<td></td>
</tr>
<tr>
<td>mean–min</td>
<td>qreg linear error</td>
<td></td>
</tr>
<tr>
<td>sharpness</td>
<td>qreg quadratic error</td>
<td>qreg contour centroid</td>
</tr>
</tbody>
</table>

4.3. Merit Estimation from Acoustic Features

After confirming the effectiveness of 2-dimensional mapping, our concern shifted to the relationship between the merit space and the acoustic feature. What is the meaning of \( \xi \) and \( \eta \)? How can we estimate \( \xi \) and \( \eta \) from the voice signal itself?

Since the size of our database is not large enough for two-stage (optimal mapping and merit estimation) fully open condition experiments, the experiments presented here were conducted under a semi-open condition. The mapping itself shown in Figure 2 was obtained using all the data. However, after the correct merit labels \( \{ \xi \} \) and \( \{ \eta \} \) were obtained for all the data, the estimation task of the merits from the acoustic features was investigated using the 10-fold cross validation (CV).

Table 3: Comparison of optimal mapping efficiency.

<table>
<thead>
<tr>
<th></th>
<th>1-d</th>
<th>2-d</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_C )</td>
<td>1068</td>
<td>1143</td>
</tr>
<tr>
<td>( N_D )</td>
<td>232</td>
<td>186</td>
</tr>
<tr>
<td>( N_T )</td>
<td>88</td>
<td>59</td>
</tr>
<tr>
<td>( \tau_b )</td>
<td>0.622</td>
<td>0.705</td>
</tr>
</tbody>
</table>
As mentioned before, we started the experiments using all of the 325 features. SMOreg estimator with the second-order polynomial kernel was used. BSS was carried out using the criteria of the average CV correlation between the true merit obtained by the optimal mapping and the estimated merit from the acoustic feature. Figure 4 shows the results of BSS. Even though the estimation using all features was very poor, the correlation value increases as the number of used features decreases. In these experiments, in addition to the estimation of $\xi$ and $\eta$, their sums and differences were also estimated. If their sum and difference are estimated from the acoustic feature, it is straightforward to estimate $\xi - \eta$, $\xi + \eta$, and $\eta$. Although it is slightly difficult to estimate $\xi - \eta$, four sets of experiments showed similar tendencies, in that the highest correlation was obtained using around 50 features. In the case of Figure 4, the highest correlation values were 0.909 (76 features) for $\xi$, 0.925 (60 features) for $\eta$, 0.944 (42 features) for $\xi + \eta$, and 0.901 (77 features) for $\xi - \eta$. Similar experiments were carried out using a 1-dimensional merit space, and the largest correlation value for $\xi$ was 0.929 (60 features).

It would be interesting to see which feature had the strongest influence on the correlation. We checked the detail of BSS right after the largest correlation was obtained. The largest likelihood drop caused by a single feature removal indicates that the removed feature plays an important role, although the features are not independent of each other. The features that indicated the strongest influence on each merit value are shown in Table 4. Interestingly, the same feature appears in two places, $\eta$ and $\xi+\eta$. Overall, the peak position of spectral features seems to be important for “irasshaimase” attractiveness.

Finally, the predictability of comparison results from the acoustic feature was evaluated using $\tau_b$. The attractiveness of the voice signal for a specific listener can be calculated using the merit values and the preference angle shown in Figure 5. For each pair, the comparison result was judged as concordant, discordant or tied, and the total efficiency was calculated as $\tau_b$. The results are shown in Table 5 with the reference values obtained by the 1-dimensional experiments.

In the 1-dimensional case, although $a$ can be estimated with a high correlation, $\tau_a$ decreased severely from Table 3 to Table 5. In the 2-dimensional case, $\tau_b$ was kept relatively high. Although the difference between estimating $\xi$, $\eta$ and estimating $\xi + \eta$, $\xi - \eta$ was not large, the latter showed a slightly better value of $\tau_b = 0.653$, which is even higher than $\tau_a$ for the 1-dimensional optimal mapping (0.622).

5. Conclusions

In this paper, a multidimensional mapping scheme of voice attractiveness was proposed. Even if different listeners have different opinions on voice attractiveness, each voice can be mapped onto a specific position of the merit space, and the interlistener variety can be attributed to the preference direction associated to the listener.

The effectiveness of the proposed model was confirmed by the experiments using a real database. Some inconsistency in the comparison results can be resolved if 2-dimensional merit space is introduced. We also investigated the relationship between the merit space model and the acoustic feature. If the estimation model was prepared by machine learning, comparison results of voice pairs can be predicted. The effectiveness of the proposed model also contributed to the improvement of prediction accuracy.

Some of the findings of this paper may be applicable for a specific greeting “irasshaimase” only, so more investigation with various utterances would be needed. Increasing the size of the database is obviously another important future work.

---

2It should be noted that there might be a set of intercorrelated features that are very important as a group, but removing only one of them does not decrease the correlation very much.
6. References


