INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Personalized Quantification of Voice Attractiveness in Multidimensional Merit Space

Yasunari Obuchi

School of Media Science, Tokyo University of Technology
obuchiysnr@stf.teu.ac.jp

Abstract

Voice attractiveness is an indicator which is somehow objective and somehow 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 guess the speaker's emotion from the recorded voices. In this case, the speaker can speak more naturally. In addition, the labels obtained by the second method can also be interpreted as the listener's emotion triggered by the voice signal. If we ask the listeners to answer how attractive the voice is, the labels obtained can be used as the training corpus for the voice attractiveness estimator.

In this paper, we focus on the problem of how we can obtain reliable labels of voice attractiveness for a given set of voice signals. Since it is difficult for ordinary listeners to evaluate the attractiveness in an absolute scale [3], we adopt the paired comparison test, in which the listener listens to two voices and answers which is more attractive. The results of the paired comparison are analyzed in a probabilistic model, in which the global consistency, personal consistency within the listener, and uncertainty are discussed separately. Even in a probabilistic model, the existence of listener pairs with opposing opinions leads to a low likelihood. To provide a better-fit model for such situations, a multidimensional merit space is introduced in which the characteristics of voices and the preference of listeners are handled in a unified model. We also analyze the relationship between the merit space model and the acoustic feature, and prove the effectiveness of the proposed model by showing the higher predictability of the comparison result.

2. Analysis of Paired Comparison

2.1. Universal Attractiveness Model

Subjective evaluation of voice signals has been used in many situations. A common example is the quality assessment of synthesized speech. 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,

$$L = \sum_i \sum_j w_{ij} \log f(d_{ij}) \tag{1}$$

where w_{ij} is the number of times the voice i was preferred to the voice j, $d_{ij} = a_i - a_j$ is the difference of the attractiveness between the voices i and j (a_i is the attractiveness of the voice i), and $f(d_{ij})$ is the probability by which i is preferred to j. In the Bradley-Terry model [8], the probability is defined as though two voices are competing for the shared resource as follows.

$$f(d_{ij}) = \frac{\mathrm{e}^a_i}{\mathrm{e}^{a_i} + \mathrm{e}^{a_j}} = \frac{1}{1 + \mathrm{e}^{-d_{ij}}} \tag{2}$$

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,

$$f(d_{ij}) = \frac{1}{2}(1 + \mathrm{erf}(d_{ij})) \tag{3}$$

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.

http://dx.doi.org/10.21437/Interspeech.2017-130

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,

$$L = \sum_i \sum_j \sum_k w_{ijk} \log f(d_{ijk}) \tag{4}$$

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 $d_{ijk} = a_{ik} - a_{jk}$ does not change, where the attractiveness for the listener k is defined as,

$$a_{ik} = \mathbf{p}_k \cdot \mathbf{m}_i \tag{5}$$

where $\mathbf{p}_k$ ($|\mathbf{p}_k| = 1$) is the preference vector for the listener k and $\mathbf{m}_i$ 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, $\mathbf{p}_k = (\cos\theta_k, \sin\theta_k)^T$ where θ_k is called the **preference angle**, and $\mathbf{m}_i = (\xi_i, \eta_i)^T$ where ξ_i and η_i are called **merit values**. The gradient can be obtained by differentiating L in terms of these three variables. (Note that $f(d_{ijk}) + f(d_{jik}) = 1$)

$$\frac{\partial L}{\partial \xi_i} = \sum_{j \neq i} \sum_k f'(d_{ijk}) \left(\frac{w_{ijk}}{f(d_{ijk})} - \frac{w_{jik}}{1 - f(d_{ijk})} \right) \cos\theta_k \tag{6}$$

$$\frac{\partial L}{\partial \eta_i} = \sum_{j \neq i} \sum_k f'(d_{ijk}) \left(\frac{w_{ijk}}{f(d_{ijk})} - \frac{w_{jik}}{1 - f(d_{ijk})} \right) \sin\theta_k \tag{7}$$

$$\frac{\partial L}{\partial \theta_k} = \sum_i \sum_{j \neq i} f'(d_{ijk}) \left(\frac{w_{ijk}}{f(d_{ijk})} - \frac{w_{jik}}{1 - f(d_{ijk})} \right) r_{ijk} \tag{8}$$

$$d_{ijk} = (\xi_i - \xi_j) \cos\theta_k + (\eta_i - \eta_j) \sin\theta_k \tag{9}$$

$$r_{ijk} = (\eta_i - \eta_j) \cos\theta_k - (\xi_i - \xi_j) \sin\theta_k \tag{10}$$

If we have a limited number of comparison results, it is reasonable to put constraints on ξ, η, and θ to remain in a given range. For simplicity, we assume that the merit values are constrained in [0, 1]. We also assume that ξ and η have nonnegative influence to the attractiveness for all listeners. Accordingly, the range of θ 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

$$f'(d_{ijk}) = \frac{1}{\sqrt{\pi}} \mathrm{e}^{-d_{ijk}^2} \tag{11}$$

```
repeat
  initialize $\{\xi_i\}$, $\{\eta_i\}$, $\{\theta_k\}$ randomly
  repeat
    for all i, k do
      calculate $\frac{\partial L}{\partial \xi_i}$, $\frac{\partial L}{\partial \xi_i}$, $\frac{\partial L}{\partial \xi_i}$ using (6),(7),(8)
    end for
    for all i, k do
      $\xi_i \leftarrow \xi_i + \frac{\partial L}{\partial \xi_i} s$
      $\xi_i \leftarrow \max(\min(\xi_i, 1), 0)$
      $\eta_i \leftarrow \eta_i + \frac{\partial L}{\partial \eta_i} s$
      $\eta_i \leftarrow \max(\min(\eta_i, 1), 0)$
      $\theta_k \leftarrow \theta_k + \frac{\partial L}{\partial \theta_k} s$
      $\theta_k \leftarrow \max(\min(\theta_k, \frac{\pi}{2}), 0)$
    end for
  until converge
  calculate $L$ using (4) and store $\{\xi_i\}$, $\{\eta_i\}$, $\{\theta_k\}$, $L$
until $N$ times
return $\{\xi_i\}$, $\{\eta_i\}$, $\{\theta_k\}$ that yielded the largest $L$
```

Figure 1: *Psudocode for log likelihood maximization.*

Table 1: *List of low level descriptors*

Energy/Pitch	Spectral	
RMS energy	max position	skewness
Log energy	min position	kurtosis
F0	centroid	slope
voicing prob.	entropy	harmonicity
	variance	sharpness

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, voiacce input is divided into 25 ms overlapping frames with 10 ms frame shift, and various features (called low level descriptors: LLDs) are extrted. 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).

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

Extremes	Regression	Segment
max	linreg slope	number of seg
min	linreg offset	seglen mean
range	linreg linear error	seglen max
max position	linreg quadratic error	seglen min
min position	qreg coef 1	seglen std. dev.
mean	qreg coef 2	
max−mean	qreg coef 3	
mean−min	qreg linear error	
sharpness	qreg quadratic error	
	qreg contour centroid	

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 $\mathbf{p_k}$ and $\mathbf{m_i}$. 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_i. The results are shown in Figure 3.

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

$$\tau_b = \frac{N_C - N_D}{\sqrt{N_C + N_D + N_T}\sqrt{N_C + N_D}}. \quad (12)$$

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 exhaustive, τ_b was calculated with 1,388 compared pairs only. From its definition, $-1 \leq \tau_b \leq 1$ was always satisfied, and $\tau_b = 1$ means that all comparison results were correct.

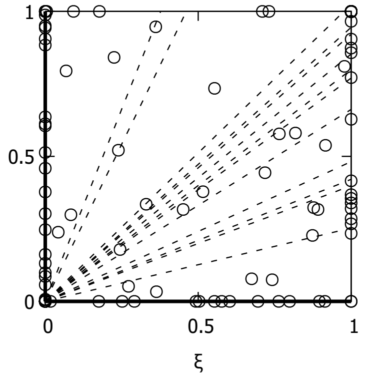

Figure 2: *Optimal mapping of 2-dimensional merit space. The voices are represented by circles. There are two listeners with $\theta_k = 0$, represented by the x-axis. There are three listeners with $\theta_k = \pi/2$, represented by the y-axis. Other listeners are represented by the dashed lines, each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

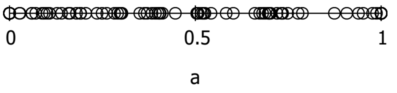

Figure 3: *Optimal mapping of 1-dimensional merit space. The variable a is the 1-dimensional merit value. There are 26 voices mapped to $a = 0$, and 22 voices mapped to $a = 1$.*

Table 3 shows the comparison of 1- and 2- dimensional optimal mapping in terms of τ_b. It can be observed that τ_b was increased by 0.083 points by introducing 2-dimensional mapping.

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 ξ and η? How can we estimate ξ and η 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 $\{\xi_i\}$ and $\{\eta_i\}$ 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.*

	1-d	2-d
N_C	1068	1143
N_D	232	186
N_T	88	59
τ_b	0.622	0.705

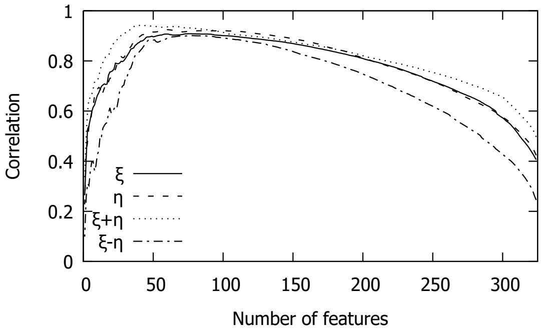

Figure 4: *Results of backward stepwise selection.*

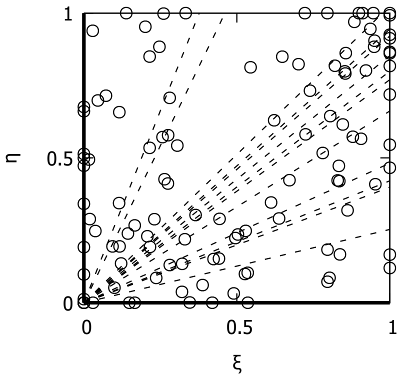

Figure 5: *Estimated mapping of 2-dimensional merit space. The voices are represented by circles. Dashed lines representing preference angles are copied from Fig. 2.*

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 ξ and η, their sums and differences were also estimated. If their sum and difference are estimated from the acoustic feature, it is straightforward to estimate ξ and η themselves. 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 ξ, 0.925 (60 features) for η, 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 a 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, η and $\xi+\eta$. Overall, the peak position of spectral features seems to be important for “irasshaimase” attractiveness[2].

Table 4: *Most influential feature for each merit value.*

	LLD	Functional
ξ	spectral entropy	max
η	spectral variance	max position
$\xi + \eta$	spectral variance	max position
$\xi - \eta$	spectral variance	min position

Table 5: *Comparison of estimated mapping efficiency.*

	1-d	2-d (ξ, η)	2-d $(\xi+\eta, \xi-\eta)$
N_C	1084	1142	1146
N_D	300	246	240
N_T	4	0	2
τ_b	0.566	0.646	0.653

Finally, the predictability of comparison results from the acoustic feature was evaluated using τ_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 τ_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, τ_b decreased severely from Table 3 to Table 5. In the 2-dimensional case, τ_b was kept relatively high. Although the difference between estimating ξ, η 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 τ_b 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 inter-listener 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.

[2] It 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

[1] M. Valstar et al., “AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge,” *Proc. 6th International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, 2016.

[2] B. Schuller, et al., “The INTERSPEECH 2016 Computational Paralinguistic Challenge: Deception, Sincerity & Native Language,” *Proc. INTERSPEECH 2016*, San Francisco, CA, USA, 2016.

[3] N. B. Shah, et al., “When is it Better to Compare than to Score?,” arXiv:1406.6618v1 [stat.ML], 2014.

[4] F. Ribeiro, et al., “CROWDMOS: An Approach for Crowdsourcing Mean Opinion Score Studies,” *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2011.

[5] H. Zen, et al., “Hidden Semi-Markov Model Based Speech Synthesis,” *Proc. INTERSPEECH 2004*, Jeju Island, Korea, 2004.

[6] J. Yamagishi, et al., “Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis,” *IEICE Trans. Information and Systems*, Vol.E88-D, No.3, pp.502-509, 2005.

[7] M. Cattelan, C. Varin, and D. Firth, “Dynamic Bradley-Terry Modelling of Sports Tournaments,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol.62, No.1, pp.135-150, 2013.

[8] R. A. Bradley and M. E. Terry, “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons,” *Biometrika*, Vol.39, No.3/4, pp.324-345, 1952.

[9] F. Mosteller, “Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations,” *Psychometrika*, Vol.16, No.1, pp.3-9, 1951.

[10] Y. Fujimoto, H. Hino, and N. Murata, “Item-User Preference Mapping with Mixture Models - Data Visualization for Item Preference,” *Proc. International Conference on Knowledge Discovery and Information Retrieval*, pp.105-111, 2009, doi: 10.5220/0002274001050111

[11] D. Causeur and F. Husson, “A 2-dimensional Extension of the Bradley-Terry Model for Paired Comparisons,” *Journal of Statistical Planning and Inference*, Vol.135, pp.245-259, 2005.

[12] F. Eyben, et al., “Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor,” *Proc. ACM Multimedia (MM)*, Barcelona, Spain, 2013, doi:10.1145/2502081.2502224

[13] A. Lee and T. Kawahara, “Recent Development of Open-Source Speech Recognition Engine Julius,” *Proc. APSIPA Annual Summit and Conference*, Sapporo, Japan, 2009.

[14] M. Hall, et al., “The WEKA Data Mining Software: an Update,” *SIGKDD Explorations*, Vol.11, No.1, pp.10-18, 2009.

[15] S. K. Shevade, et al., “Improvements to the SMO Algorithm for SVM Regression,” *IEEE Trans. Neural Networks*, Vol.11, No.5, pp.1188-1193, 2000.