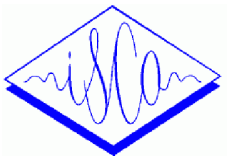
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Predicting Likability of Speakers with Gaussian Processes

Dingchao Lu, *Fei Sha**

*Department of Computer Science, U. of Southern California, Los Angeles, CA 91007, USA
dingchal@usc.edu, feisha@usc.edu

Abstract

In this paper, we study the problem of predicting likability of speakers based on their voices. We use the data provided by the Likability Sub-Challenge of InterSpeech 2012 Speaker Trait Challenge. We explore Gaussian Processes (GP) based learning techniques for classification and regression. In particular, we propose a novel algorithm that greedily finds a sparse yet informative subset of features. We also show how the covariance functions learned for GP models can be used to derive new features for prediction. The best system that we have developed integrates those techniques and achieves 60.1% unweighted accuracy on the Likability Sub-Challenge (1.86% relative improvement over the provided baseline). The proposed approach also works well on the Pathology Sub-Challenge, achieving an accuracy of 73.7% (6.92% relative improvement).

Index Terms: likability of voice, Gaussian Processes, sparse models, intelligibility of voice

1. Introduction

Imagine a simple program that can analyze your voice and tell you how likable or pleasant you sound. Even better would be if the program can tell you what acoustic qualities you can improve on. Such a likability algorithm would be widely appreciated by anyone who needs to communicate more effectively, from makers of turn-by-turn GPS, to political candidates out to rally the masses, to the average Joe getting ready for a blind-date.

Despite the pervasiveness of voices in our daily lives, determining the likability of speakers based on their voices, as opposed to the content of their speech, is not a trivial task. Previous human behavior studies have shown that women prefer low-pitched voices, an indicator of high testosterone levels, in men [1], while men strongly agreed that high frequency female voices were more attractive [2]. However, state-of-the-art computer algorithms are far from being able to accurately quantify likability in voices.

A recent study that used a database of recorded German telephone speech achieved a best unweighted accuracy of 67.6% with random forests when predicting binary likability labels [3]. This database was transformed into the Speaker Likability Database (SLD) and a standard acoustic feature set was extracted for the Likability Sub-Challenge of Interspeech 2012 Speaker Trait Challenge [4]. The official scoring metric is unweighted accuracy (UA) on test data and the published baselines on SLD are 55.9% for support vector machine (SVM) and 59.0% for random forests.

In this work, we propose new machine learning algorithms for the prediction task. We focus on the technique of Gaussian Processes (GP) for classification and regression due to its flexibility in modeling data. In particular, many parameters in the learning algorithm can be automatically inferred.

We propose a novel algorithm that greedily selects a sparse yet informative subset of features. We show that parsimonious models, which do not use all features, attain higher accuracies. We show how to improve performance on the test data using semi-supervised learning techniques where both training, development and test data are embedded into a new feature space.

We have also shown empirically that there is a clear distinction in the predictability between male and female voices. Specifically, male-specific models often have higher accuracies.

The best system that we have developed achieves 60.1% UA on likability (1.1% absolute and 1.86% relative improvement over the published baseline of 59.0%). Furthermore, we applied the proposed techniques for the Pathology Sub-Challenge to predict the intelligibility of voices of Dutch patients undergoing chemo-radiation treatment [4]. The proposed method works well and outperforms the published baseline of 68.9% by a significant margin and attains an accuracy of 73.7%.

2. Approach

In this section, we briefly review learning algorithms that we have developed for the prediction task. While our own baselines use popular techniques such as SVM, our best performing systems use extensively Gaussian Processes (GP) for classification and regression [5].

To overcome the challenge of high-dimensional features, we extend the standard GP so that it can be used to construct models using a sparse (sub)set of features. We also show how the covariance functions learned by the GP models can be used to derive new features to enable

semi-supervised learning. In what follows, we describe both the standard technique and our extensions.

2.1. Gaussian Processes

Gaussian Processes (GP) is a nonparametric Bayesian framework for modeling data. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ denote N training samples, while $\boldsymbol{x}_i \in \mathbb{R}^D$ is a D-dimensional feature and y_i is the corresponding label, which can be either categorical for classification or continuous for regression. The training data defines a prior on the functional form between $\boldsymbol{x}$ and y.

GP for Regression (GPR) For regression problems, the prior is on a continuous function $f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x}'))$. The mean function $m(\boldsymbol{x})$ is often conveniently assumed to be zero. The scalar kernel (or covariance) function $K(\boldsymbol{x}, \boldsymbol{x}')$ encodes how two different features $\boldsymbol{x}$ and $\boldsymbol{x}'$ are correlated. For a finite set of features $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, the prior implies that the joint distribution of the function values on those features is a multivariate Gaussian. The covariance matrix of the Gaussian is $\boldsymbol{K}$ whose ij-th element is $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Common choices of the kernel function include Gaussian RBF kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2/\sigma^2\}$, and linear kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$. The parameters of these functions are then inferred from the training data.

Under this prior, the label y (ie, the function value) of a new feature vector $\boldsymbol{x}$, is governed by the predictive distribution $p(y|\boldsymbol{x}, \mathcal{D})$, which is a univariate Gaussian. The mean and the variance of this Gaussian can be computed efficiently using standard procedures for marginalization and conditioning of multivariate Gaussians.

GP for Classification (GPC) GP for classification is substantially more computationally intensive. We explain the key concepts and leave details to references [5].

In the setting of classification, GP assumes that the label y is a squashed value of another latent function $f(\boldsymbol{x})$ whose prior is given by the training data, as in regression. For binary classification, the "squashing" maps the range of $f(\boldsymbol{x})$, which is often the whole real axis, to the interval $[0\,1]$. There are two common choices for the squashing function: the logistic function and the erf function, which are often referred as likelihood functions in GP literature. In this work, we have used the logistic function.

Advantages of GP By flexibly specifying functional forms of the mean and covariance functions, GP includes a very rich set of models (including nonlinear ones). Particularly, multiple (base) covariance functions can be used to compose complex ones:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_m \alpha_m K_m(\boldsymbol{x}_i, \boldsymbol{x}_j) \qquad (1)$$

where each base function $K_m(\boldsymbol{x}_i, \boldsymbol{x}_j)$ can be defined independently. In this work, we have used different groups of features to define base functions, see section 2.2. The combination coefficients α_m, as well as each base function's own parameters can be *automatically* inferred [5].

2.2. Sparse Gaussian Processes (S-GP)

The high-dimensional feature vectors in the speaker trait prediction tasks present a serious challenge for statistical modeling, as there is only a small number of labeled examples. It is reasonable to hypothesize that some of the features are useful while others could be less effective or redundant. Thus, identifying what features are most useful for predicting the speaker trait is both insightful to the understanding of what acoustic qualities reflect the trait and also beneficial in preventing overfitting.

To this end, we extend the original framework of GP and propose a sparse Gaussian processes (S-GP) method, similar in spirit to the technique of orthogonal matching pursuit (OMP) used in signal processing [6].

Specifically, we leverage the fact that the 6125 feature dimensions from the dataset naturally form into 13 groups, according to how they are derived, cf. Table 1. Concretely, the method consists of the following key steps: i) *Initialization*. The selected feature group is initialized to be an empty set. ii) *Greedy selection*. The selected feature group is augmented by adding to it the optimal one from the feature groups that have not been selected. The optimal one is defined as the feature group that, when combined with the selected ones, achieves the highest accuracy. The combination takes the form of eq. (1). iii) *Iteration*. The procedure iterates back to step ii) until the accuracy is no longer improved.

In practice, this S-GP procedure performs well in practice by using only a few feature groups.

2.3. Semi-supervised learning

The kernel function eq. (1), once learned, can be used to derive new features for data. Specifically, we use the learned kernel in kernelized PCA (KPCA) to compute a *joint* embedding of the training, development and test instances. In our work, we compute the kernel matrix and identify its eigenvectors and then project each instance into the eigenvectors to obtain the new features.

Once the new features are learned, we use those corresponding to the training and development to build a second GP model and to predict on the test data. This approach has a flavor of semi-supervised learning as the new features are computed using labeled information from the training and development, as well as the unlabeled test data. Empirically, this stage yields a noticeable improvement on the performance.

3. Experimental Results

In this section, we report results of applying algorithms described in section 2 to the Likability and Pathology challenges. We start by summarizing the general setup

Table 1: 13 features groups: names of the low-level descriptors (LLD), # of feature dimensions (D), and individual average 10-fold CV accuracies for the Likability (L-UA) and the Pathology (P-UA)

LLD Names	Grp #	D	L-UA	P-UA
Energy Related Groups				
Sum of aud. spec. (loudness)	1	96	58.0	67.5
Sum of RASTA aud. spec.	2	96	54.5	65.5
RMS energy	4	96	60.1	66.7
Zero-crossing rate	5	96	55.3	67.3
Spectral Related Groups				
RASTA-style aud. spec.	3	2496	57.1	72.3
MFCC 1-14	7	1344	59.5	77.2
Spectral	6	1344	59.3	73.9
Voicing Related Groups				
F0	8	97	56.0	67.0
Voicing	9	92	57.7	68.0
Jitter local	10	92	55.6	65.0
Jitter delta	11	92	57.4	65.7
Shimmer local	12	92	55.6	64.3
Log Harmonic-to-Noise Ratio	13	92	56.3	66.8

of our empirical studies, followed by describing results of our baseline systems on the *development* data. These baseline systems are improved by the following three extensions: i) modeling male and female voices separately (only for likability), ii) using a sparse subset of features to build statistical models, and iii) creating new features with Kernel PCA that incorporates unlabeled test data. We observe improvement in accuracies over published baselines for both challenges. We report those results too.

3.1. Setup

For Likability, we use the Speaker Likability Database (SLD) and keep the same split of training, development and test [4]. In addition to binary labels, the dataset also provides evaluator weighted estimator (EWE), a continuous score of likability. We also experimented regressing on this variable and then thresholding continuous predictions to obtain binary labels. The threshold is set at 0.11.

The database provides an extensive acoustic feature set extracted with openSMILE. We grouped the 6125-dimensional feature vectors into 13 groups, displayed in Table 1. We used these groups to define different base Gaussian RBF kernels, as described in section 2. We z-scored each feature dimension. We have used the publicly available implementation of Gaussian Processes techniques [5]. As a reference point, the prediction accuracies of individual base kernels are also reported in Table 1. The unweighted accuracies (UA)s are computed with GPR using 10-fold *cross-validation* on combined training and development data. In general, MFCC appears to be very informative.

For Pathology, the dataset has identical structures as the one for the Likability except one subtle difference. The EWE for the Pathology is asymmetrically distributed on a 1 to 7 scale with the INTELLIGIBLE class above 5.75 and NON-INTELLIGIBLE class below. Before using the EWE for regression, we linearly scale up the INTELLIGIBLE EWEs so they are symmetric about 5.75 with the NON-INTELLIGIBLE EWEs. Table 1 shows individual feature group accuracies using GPR.

Table 2: Prediction accuracies of our baselines on development for the Likability (L-UA) and Pathology (P-UA)

Method	L-UA	P-UA
Linear SVM w/ all features	58.5	62.0
GPC w/ 1 Gaussian kernel	59.6	62.5
GPR w/ 1 Gaussian kernel	60.6	62.8
GPC w/ combined 13 Gaussian kernels	57.5	64.2
GPR w/ combined 13 Gaussian kernels	61.9	63.3

3.2. Baselines

Table 2 reports our own baseline accuracies on the *development* data. Our results closely match the baselines published by the Challenge organizers [4].

Gaussian Processes classification (GPC), described in section 2, yields a slightly better accuracy than the baseline linear SVM for both Likability and Pathology. This is likely due to the nonlinear classification, as we have used Gaussian RBF kernel for the kernel functions in GP. Note that GPR improves over GPC slightly.

We improve further by defining 13 Gaussian RBF kernels, one for each feature group (cf. Table 1) and combine them as in eq. (1). This adds extra modeling flexibility as each base kernel has its own parameters to be tuned and the importance of different groups can be weighted by the combining weights.

3.3. Gender difference in predicting likability

Since the Likability dataset contains a balanced number of male and female voices, one naturally wonders whether it is equally difficult to predict likability of each gender. For example, earlier studies suggest that male voices, when judged by female evaluators, tend to receive more consistent assessments [1]. This seems to hint that the likability of male voices might be easier to determine.

Thus, we build and evaluate gender-specific models of likability, using corresponding data in training and development respectively. GP models combining 13 base kernels (as in Table 2) perform the best, so we report results of those models in Table 3. The male-specific model is more accurate. Moreover, regression is more accurate for males while classification is better for females.

The test data does *not* have gender information about the speaker. However, we find that we can predict gender on development data with 95% accuracy using a simple GPC. Thus when we use gender-specific models on test data, we first predict the gender then apply the appropriate gender model to predict likability.

Due to unbalanced genders in the Pathology, we did not perform the gender-specific modeling and analysis.

Table 3: Accuracies of gender-specific models using 13 combined Gaussian kernels

Gender	Method	UA on Development (%)
Female	GPC	**64.2**
Female	GPR	60.7
Male	GPC	65.3
Male	GPR	**70.6**

Table 4: Average CV accuracy of non-sparse and sparse models, and selected feature groups for the Likability

Method	Non-sparse model	Sparse model	Selected groups
GPC (Females)	61.0	63.7	11, 13, 4, 5
GPR (Females)	56.0	62.8	11, 4
GPC (Males)	61.0	66.2	1, 9,3, 2, 12, 7, 8
GPR (Males)	64.8	70.2	4, 13, 8, 10, 3, 6, 9, 2

3.4. Sparse models

We employ the S-GP method described in section 2.2 to examine which feature groups are among the most effective in predicting the speaker traits. To prevent overfitting on the development data, we combine the training and the development as a new and larger training data set. When we add a candidate feature group, we apply cross-validation (CV) and compute the average UA of the resulting models across the different folds of CV. The candidate feature with the highest average UA is then added to the set of selected feature groups. From here on, unless otherwise noted, all accuracies are based on cross-validation (10-fold for Likability and 5-fold for Pathology). Also, all models are retrained with the combined train and development sets before predicting on test.

The second column of Table 4 shows the average CV UA's of the 4 gender-specific models from Table 3. Note that the sparse models outperform non-sparse ones.

We observe that in general, female-specific models use less feature groups. Considering that prediction accuracies on female voices are typically lower than those on male voices, it is quite likely that existing features are inadequate in capturing the complexity of female voices.

Additionally, it seems that male and female models use a rather disjoint set of features. For classification, there are no overlapping feature groups. For regression, only feature group #4 (RMS energy) is shared.

3.5. Semi-supervised learning

We applied the semi-supervised learning method described in section 2.3 to learn a new feature representation *after* learning the kernel function in the GP models. This results in improved performance, as shown in Table 5.

In the table, we combined the best female model (S-GPC) and the best male model (S-GPR) as described in Table 4 for prediction on the test set. It modestly outperforms the published baseline of 59.8%.

We also built a sparse GPR that is gender-independent and predicted on the test data, which yielded a worse accuracy of 58.3%. However, using the semi-supervised learning (S-GPR+KPCA), we obtained a noticeably improved accuracy of 60.1%, which is 1.1% better in absolute than the published baseline. (Due to time constraints, we did not build a system that uses gender-specific sparse models with semi-supervised learning.)

Table 5: 10-fold CV average accuracy and test accuracy for Likability. Baseline for test is 59.0%.

Method	Selected Groups	CV UA	Test UA
S-GPC Females + S-GPR Males	11, 13, 4, 5 + 4, 13, 8, 10, 3, 6, 9, 2	63.7 + 70.2	59.8
S-GPR	4, 8, 3, 10, 13, 7, 1, 11	64.52	58.3
S-GPR + KPCA	4, 8, 3, 10, 13, 7, 1, 11	62.11	**60.1**

Table 6: 5-fold CV average accuracy and test accuracy for Pathology. Baseline for test is 68.9%.

Method	Selected Groups	CV UA	Test UA
GPC	All 13	78.5	69.9
S-GPC	7, 6, 5, 1, 9, 12, 13	80.1	68.7
GPR	All 13	75.7	72.0
S-GPR	7, 4, 13, 6, 1, 5, 8	77.4	72.1
S-GPR + KPCA	7, 4, 13, 6, 1, 5, 8	77.6	**73.7**

3.6. Results on the Pathology Sub-challenge

As show in table 6, our best result is 73.7%, significantly better than the baseline 68.9%. (Detailed analysis will be reported in a longer version of this paper.)

4. Discussion

We have applied and extended Gaussian Processes to both likability and intelligibility prediction. Our key observation is that sparse models are effective in controlling overfitting due to the high-dimensional feature vectors. The proposed methods are robust – works well on both challenges. In future work, we plan to extend them with other robust modeling techniques.

Acknowledgments This work is supported by Rose Hills Foundation (D.L.) and IARPA BABEL (F.S. and D.L.)

5. References

[1] L. Bruckert, J. Liénard, A. Lacroix, M. Kreutzer, and G. Leboucher, "Women use voice parameters to assess men's characteristics," *Proc. of the Roy. Soc. B: Biol. Sci.*, vol. 273, no. 1582, 2006.

[2] S. Collins and C. Missing, "Vocal and visual attractiveness are related in women," *Animal Behaviour*, vol. 65, no. 5, 2003.

[3] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, "Would you buy a car from me?-on the likability of telephone voices," in *InterSpeech*, 2011.

[4] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge," in *Proceedings INTERSPEECH*, 2012.

[5] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[6] Y. C. Pati, R. Rezaiifar, Y. C. P. R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. of Asilomar Conf. on Sig., Sys., & Comps.*, 1993.