An Automatic Voice Pleasantness Classification System based on Prosodic and Acoustic Patterns of Voice Preference

Luis Coelho[1,3], *Daniela Braga*[2,3], *Miguel Sales-Dias*[2,3], *Carmen Garcia-Mateo*[4]

[1]Polytechnic Institute of Porto, ESEIG, Portugal
[2]Microsoft Language Development Center, Microsoft, Portugal
[3]ADETTI - ISCTE, IUL, Portugal
[4]Signal and Communications Theory Department, University of Vigo, Spain
lcoelho@eu.ipp.pt, {dbraga, miguel.dias}@microsoft.com, carmen.garcia@uvigo.es

Abstract

In the last few years the number of systems and devices that use voice based interaction has grown significantly. For a continued use of these systems the interface must be reliable and pleasant in order to provide an optimal user experience. However there are currently very few studies that try to evaluate how good is a voice when the application is a speech based interface. In this paper we present a new automatic voice pleasantness classification system based on prosodic and acoustic patterns of voice preference. Our study is based on a multi-language database composed by female voices. In the objective performance evaluation the system achieved a 7.3% error rate.

Index Terms: Speech analysis, Speech synthesis, Human voice

1. Introduction

In recent years the speech synthesis technology has been widely improved and has reached a maturity level that leveraged their inclusion on systems and devices for daily use. GPSs, PDAs, e-learning systems or reading assistants are just a few examples. Most mainstream operating systems, for desktop and mobile devices, are also providing text to speech (TTS) systems that can work alone or can be easily integrated with other applications. The quality of these systems can be very good, with fluent speech, high intelligibility rates and even emotions in some cases. However there are still users who don't feel completely satisfied and try several voices (when available) with the purpose of finding the one that best suits their needs and personal tastes. These additional demands are related with subjective speech characteristics and, as far as the authors' knowledge, there are no studies that embrace the user's reaction to a given speech utterance or that evaluate the suitability of a voice for a given task. With the purpose of building new TTS voice fonts we explored several subjective voice features and speech interaction use cases. For this paper we have focused specifically on the concept of pleasantness according with the definitions found in [1] where ”Pleasantness is the feeling caused by agreeable stimuli.” and in [2] where pleasantness is what gives ”a sense of happy satisfaction or enjoyment”. This concept is distinct from the concept of attractiveness (”appealing to the senses”, ”sexually alluring” [2]) which was also evaluated but is out of the scope of this work. The concept of voice pleasantness, in a context of very frequent interaction, may be subject to cultural and individual variations but the obtained results pointed to several inter-cultural patterns and evidenced a substantial agreement among listeners preferences.

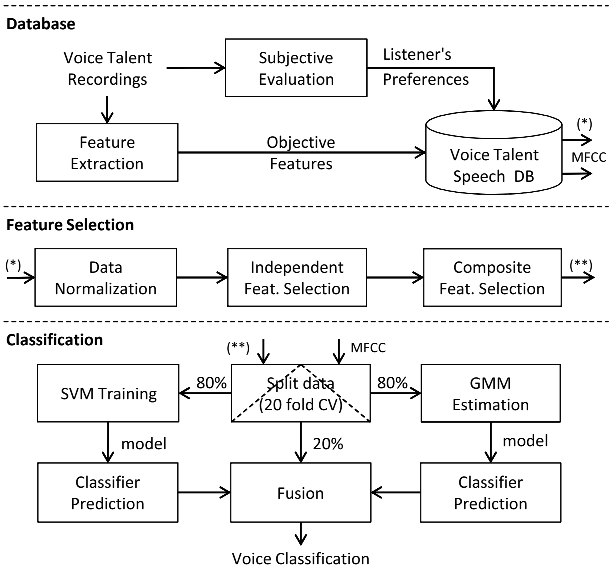

Figure 1: *System's architecture*

In this paper our goal is to objectively assess a voice and to demonstrate that it is possible to build an automatic pleasantness classification system using prosodic and acoustic patterns of voice preference. Our work is organized as follows: in the next section we present a description of the used methodology starting with an overview of the system's architecture which will then be detailed in the ensuing sub-sections. We will thoroughly cover the construction of our database, the followed pipeline for feature selection and the classifier development and tuning. Finally we will present the main results and conclusions as well as some envisioned developments.

2. Methodology

The methodology that was used for the development of the system is depicted in figure 1 with function blocks representing the main tasks sets.

2.1. Database

The database that supported the development of this work is exclusively composed by female voices. These voices belong to professional speakers and were recorded during voice talent selection processes, in the framework of for the development of new TTS systems. The related methodology as well as the quality requirements that were imposed during those processes have been previously published [3] and for this work we will only mention the relevant features. Our database covers 6 languages (Catalan, Danish, Finnish, Portuguese, Portuguese (Brazil) and Spanish) and on a previous study [4] we identified inter language trends that allowed us to homogenously consider the whole record set (for example we observed that voices whose pitch values fell between a given range usually received more listeners' preferences than others). For each language we have recordings for 5 different speakers and each one recorded around 3 minutes of speech from a common language dependent script containing phonetically and prosodically rich sentences that allowed the expression of emotions. For every language we conducted a survey where each utterance was rated according to pleasantness using a 5 points scale. The survey received an average of 60 responses per language, from male and female listeners, native and non-native speakers, in an age range from 23 to 60 years old. The database records were divided in two classes, one composed by the two best classified voices from each language and other with the remaining voices.

2.2. Features

Since we had no prior knowledge about the features that could efficiently contribute to define pleasantness we decided to build an initial prototype vector using scientifically proved features commonly used in areas that seek similar objectives, such as speech/speaker recognition, emotion recognition and clinical voice analysis. Additionally we considered the recommendations of Wolf [5] who advocates that the variables should occur naturally and frequently in normal speech, be easily measurable, have high variability between speakers, be consistent for each speaker, not change over time or be affected by the speaker's health. Our prototype vector encompassed a broad range of signal aspects, covering intra- and inter-period characteristics, time and frequency domains contents and several statistics that could complement the raw information. It was organized in four groups: acoustic features, signal features, periodicity features and phonation speed features.

In the first group we considered the fundamental frequency (f_0) envelope and its first (Δf_0) and second derivatives ($\Delta\Delta f_0$). From these we calculated four first order statistics, namely average (Av), standard deviation (Std), skewness (Sk) and kurtosis (Kt), and extracted the maximum (Max) and minimum (Min) values of the envelope (minimum was obtained excluding zeros). For four vowels, common to all the languages (in the phonetic sense), we have extracted the first four formants and their related bandwidth ($[\text{V}]_i$ represents the frequency of formant i for vowel [V]) and also calculated the four above mentioned first order statistics, maximum and minimum. The second group included the instantaneous power (P) obtained by following a similar procedure to the one described for f_0. A possible voice quality factor is the stability and cross period coherence of the signal in voiced sounds. To address this hypothesis we have included a third feature vector group, where we have considered jitter (J), shimmer (S) and harmonic to noise ratio (HNR). We calculated jitter $J(k)$ (local) in period k as the average absolute difference between consecutive periods T_i divided by the average period:

$$J(k) = \frac{\sum_{i=1}^{N} |T_i - T_{i+1}|}{\sum_{i=1}^{N} T_i} \tag{1}$$

Table 1: *Prototype Feature Vector.*

Group	Feat.	Stat.	#
Acoustic	$f_0, \Delta f_0, \Delta\Delta f_0$	Av, Std, Kt, Sk, Min, Max	18
	4xVow.: 4Fmt	Av, Std, Kt, Sk, Min, Max	96
	4xVow.: 4Bw	Av, Std, Kt, Sk, Min, Max	96
Signal	P, P', P"	Av, Std, Kt, Sk, Min, Max	18
Periodicity	Jitter	-	4
	Shimmer	-	6
	HNR, HNRdb	-	2
Phonation Speed	WR, SR, PR	-	3

The equation is based on an underlying source-filter model, where a given excitation $p(k)$, centred at instant k, composed by a sum of consecutive pulses with amplitude A_i, separated by a period T and with an admissible variation ΔT for the period i is described by:

$$p(k) = \sum_i A_i \delta(k - iT - \Delta T_i) \tag{2}$$

Besides local jitter (Jloc) we have also included other similar metrics (that can provide non-redundant information): Absolute jitter (Jabs), relative average perturbation (Jrap), period perturbation quotient and periodic difference (Jddp). The periodicity feature group also comprises shimmer whose calculation is identical to jitter but focused on the amplitudes (A_i). We accounted for six varieties of shimmer: local (Sloc), local in dB (SdB), periodic difference (Sddp) and three amplitude perturbation quotients ($S_{apqN}(k)$), calculated for 3, 5 and 11 points.

$$S_{apqN}(k) = \frac{\frac{1}{N}\sum_{i=1}^{N} \left| A_i - \frac{1}{2Q+1}\sum_{q=-Q}^{Q} A_{i+q} \right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{3}$$

In equation 3, the variable N represents the number of used pulses in the calculation and $Q = (N-1)/2$. All jitter and shimmer varieties were calculated according to Praat's [6] description. The relation between harmonic and noise signal components was also considered. We have used an harmonic to noise ratio (HNR) based on the spectral energy as shown in equation 4 and the same value in dB (HNRdB).

$$\text{HNR} = \frac{\sum_\omega |H(\omega)_{Harmonic}|^2}{\sum_\omega |H(\omega)_{Noise}|^2} \tag{4}$$

Finally, the last feature group composed by phonation speed metrics, includes the word rate (WR), as words/s, speaking rate (SR), as phonemes/s, and pause rate (PR), as the relation between pause time and total speaking time. The final prototype vector composition is shown in table 1.

Besides the described features, we have also created a speaker model using 16 Mel-frequency cepstral coefficients (MFCC) from 20ms windows with 5ms overlaps, considering 4 Gaussian mixtures.

2.3. Feature Selection

The prototype feature vector described in the previous section is composed by 243 dimensions. This number brings and increased complexity for the development of the classifier and some of the components may provide little or redundant information for the classification task. In order to identify the most discriminant feature sub-set we used the following steps. First, considering that the values on each dimension followed a Gaussian distribution, we have normalized the feature components by calculating their Z-values. This procedure centred the points in the origin and equalized the range of values preventing the domination of attributes that vary in higher numeric ranges. We have then defined as outliers all the points that exceed in value two standard deviations, in any vector dimension, and removed them (this allows to keep around 95% of the values around the mean value). A variation of the Kolmogorov-Smirnov test [7] was used to verify the initial assumption that the values on each dimension followed a normal distribution. Then a t-test, with a 90% confidence interval, was used to analyse if a given feature could be useful to discriminate between two classes. The features that have not passed these tests where discarded. Then, we have ranked the features using a metric based on the intersection of the normal cumulative distribution functions of each class, between the mean values:

$$F_{\text{rank}} = \left[k_1.erf\left(\frac{\mu_2 - \mu_1}{\sqrt{2\sigma_1^2}}\right) + k_2.erf\left(\frac{\mu_1 - \mu_2}{\sqrt{2\sigma_2^2}}\right)\right]^3 \quad (5)$$

where μ and σ represent respectively the mean and standard deviation of a given dimension for classes 1 and 2 and the Gaussian error function is defined as $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$. The variables (k_1, k_2) take the values (1,-1) or (-1,1) according with $m_1 > m_2$ or $m_2 > m_1$ respectively. This metric provides naturally range limited values which is an advantage over other metrics (like the Fischer discriminatory ratio), since we will combine it (using weights) with other variables. In figure 2 we can see the data behaviour for local jitter, one of the best ranked features. This figure shows top views of Gaussian distributions estimated for absolute rank positions (the distributions between the integer values are obtained by linear interpolation for giving a better view of the data trends). We can observe that the jitter values for the best ranked voices are concentrated in a very narrow range, while for the worst ranks, there is a wider data dispersion and a distinct mean value.

Using the ranked feature list sorted in descending order, we have built a new feature list where we took into account the inter-feature correlation. We proceeded as follows:

1. The best ranked feature f_1 is the top-ranked in the F_{rank} ranked list. The next feature f_2 is obtained by

$$f_2 = \max_j \left\{ w_1 F_{\text{rank}_j} - w_2 \rho_{f_1,j}^2 \right\}, j \neq f_1 \quad (6)$$

where F_{rank_j} is the feature's F_{rank} value, $\rho_{f_1,j}$ represents the cross-correlation between the feature in analysis and the remaining features and w_1 and w_2 are user defined weights that allow to adjust the contribution of each criteria to the overall feature ranking (a linear search pointed to $w_1 = 0.3$ and $w_2 = 0.7$).

2. The next feature is selected using an analogous formulation but now considering the average correlations with all the previously selected features:

$$f_k = \max_j \left\{ w_1 F_{\text{rank}_j} - \frac{w_2}{k-1}\sum_{r=1}^{k-1} \rho_{f_r,j}^2 \right\} \quad (7)$$

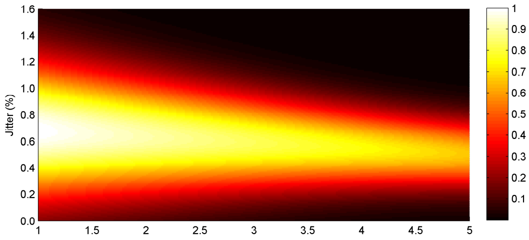

Figure 2: *Jitter distribution according to listeners' preference. Horizontal axis shows continuous candidate ranking (1 is better) and vertical axis shows jitter values in percentage.*

with $k = 3, 4, \ldots, m$, $j \neq f_r$ and $r = 1, 2, \ldots, k-1$.

This allows to obtain a feature list where the best ranked features maximize the discriminative power and minimize the redundancy. From this multi-criteria ranked feature list we have selected the 12 highest ranked features and have performed an exhaustive search for combinations of 6 features using scatter matrices [8]. This number of features was chosen to ensure a good generalization performance of the classifier, since we have a reduced number of points. As cost function for class separability measurement we used the J_3 criteria defined as:

$$J_3 = trace\left\{\mathbf{S}_w^{-1}\mathbf{S}_b\right\} \quad (8)$$

where S_w is the intraclass scatter matrix, defined as:

$$\mathbf{S}_w = \sum_{c=1}^{C} P_i \mathbf{S}_i \quad (9)$$

where P_i is the probability of each class c and $\mathbf{S}_i$ is the related covariance matrix. Still in equation 8, $\mathbf{S}_b$ represents the interclass scatter matrix,

$$\mathbf{S}_b = \sum_{i=1}^{c} P_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)^T \quad (10)$$

where $\boldsymbol{\mu}_i$ is the class mean vector and $\boldsymbol{\mu}_0$ is the mean vector considering all classes. Using an exhaustive search, we combined sets of 6 features from the 12 previously selected and retained the one that maximized the J_3 criterion. The features that composed our final optimized vector were Jrap, Δf_0Max, Δf_0Sk, ΔEAv, Sapq3and f_0Av.

2.4. Classification

We have used two distinct classification schemes: one based on Support Vector Machines (SVMs) for the optimized feature vector described in section 2.3 and another based on the Bayes decision theory for the Gaussian Mixture Models (GMM). The use of a SVM classifier, whose training relies on the optimization of a cost function, was preferred to the multilayer perceptron approach, an initial option, due to a guaranteed convergence to a global optimum and due to the simple hyperplane based model. SVMs can also provide better models in the presence of reduced data sets, which is our case. Since our data is unbalanced, the SVM problem was adapted in order to include penalties for adjusting the relative weight of each class.

The speaker model based on GMMs was trained using the Expectation-Maximization algorithm (EM)[9].

In both cases the classifier training was performed using 80% of the database records, randomly selected, and the re-

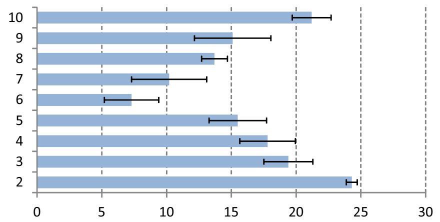

Figure 3: *SVM classification error according to feature vector dimension. Horizontal axis shows error in percentage and vertical axis shows vector dimension.*

Table 2: *Voice preference classification error for a 90% confidence interval using different kernel types with a SVM based classifier. (Two classes outlier free dataset, error within a 90% confidence interval.)*

Linear	Quadratic	RBF	Sigmoid
18.1±4.1%	12.9±1.8%	11.1±2.4%	22.7±2.9%

maining 20% were kept for testing. A 20 fold cross validation allowed to obtain statistically meaningful results.

A late fusion based on a weighted voting scheme (40%GMM+60%SVM, adjusted by calculating the relative number of true positives for each classifier against the total number of true positives) allowed to reach the final category estimation. The estimation accuracy could be further enhanced if the process is repeated using distinct models, trained with distinct data, and averaging the final decision.

3. Evaluation and Results

The feature selection process allowed us to obtain a ranked feature list from which we have extracted a fixed number of features. In figure 3 we show how the classification error varies with the number of selected features and we can observe that the best results are obtained for a 6 dimension vector (as stated before). In order to find an optimal classifier we have assessed the system's performance when varying the type of kernel function and the related parameters. Table 2 shows the best results for each kernel type considering the SVM classifier alone. We can observe that the RBF kernel performed better, but very closely followed by the quadratic polynomial kernel (despite the much longer training time required by the last). Additionally, we also evaluated how the feature selection process and how the inclusion of a second classifier (GMM) helped us to improve the results. In table 3, we can observe the improvements introduced by each block of the pipeline using as a reference the results obtained with a RBF kernel (since it performed better) and considering all the available objective features. The final system, with the components arranged as depicted in figure 1, achieved an error rate of 7.3 ± 2.2% for a 90% confidence interval.

Table 3: *Cumulative error reduction introduced by each component of the classification system. (Error within a 90% confidence interval.)*

All feat.	Independent feat. sel.	Composite feat. sel.	Composite classification
(baseline)	-14.7±3.1%	-19.2±1.9%	-23.8±2.2%

4. Conclusions

In this paper we have presented a new automatic voice talent classification system based on prosodic and acoustic patterns of voice preference. We started by introducing the motivations for this study and clearly defined the pleasantness concept that has a central role in our work. After showing an overview of the methodology we described the database on which we relied to develop our system. We have focused on the main components and provided references that fully cover it's development. Then we thoroughly explained the procedure that was followed to obtain an optimal feature vector for achieving improved results. We have started by exploring a broad range of dimensions, covering acoustic, signal, periodicity and phonation speed aspects and successively selected the best features by maximizing their class discriminatory power and by reducing the inter-feature redundancy. We have also proposed a new metric for the purpose of composite feature selection. For classification we used a combined SVM/GMM technique with a late fusion scheme. We have performed a wide evaluation of the system while varying the most important parameters and we presented how each pipeline block contributed to improve the performance. Our system achieved a final classification error rate of 7.3 ± 2.2% for a 90% confidence interval. We believe that this novel tool can be useful for reducing voice talent selection costs, therefore enhancing TTS systems, and we further think that there is a potential for the introduction and exploitation of this technology in real life applications.

5. Acknowledgements

This work was partially supported by the Spanish government and ERDF funds under the project TEC2009-14094-C04-04.

6. References

[1] C. Fellbaum, *WordNet: An Electronical Lexical Database*. Cambridge, MA: The MIT Press, 1998.

[2] *Oxford Dictionary*, 2009, [Online; accessed 13-Mar-2011]. [Online]. Available: http://oxforddictionaries.com/

[3] D. Braga, L. Coelho, F. G. V. R. Junior, and M. S. Dias, "Subjective and objective assessment of TTS voice font quality," in *Proc. of International Conference on Speech and Computers (SPECOM 2007)*, Moscow, October 2007, pp. 306–311.

[4] H.-U. Hain, O. Jokisch, and L. Coelho, "Multilingual voice analysis: Towards prosodic correlates of voice preference," in *Konferenz Elektronische Sprachsignalverarbeitung*, Dresden, 2009.

[5] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *Journal of the American Statistical Association*, vol. 51, pp. 2044–2056, 1972.

[6] P. Boersma, "Praat: doing phonetics by computer," *Glot International*, vol. 9/10, no. 5, pp. 341–345, 2001. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[7] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *J. American Statistical Association*, vol. 62, pp. 399–402, 1967.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, 1977.