Available online at www.sciencedirect.com
SciVerse ScienceDirect
Computer Speech and Language xxx (2012) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

On the development of an automatic voice pleasantness classification and intensity estimation system☆

Luis Pinto-Coelho[a,b,*], Daniela Braga[c], Miguel Sales-Dias[b,d], Carmen Garcia-Mateo[e]

[a] *Polytechnic Institute of Porto, Portugal*
[b] *Microsoft Language Development Center, Portugal*
[c] *TTS Team, Microsoft, China*
[d] *ISCTE-Lisbon University Institute, Portugal*
[e] *University of Vigo, Spain*

Received 9 May 2011; received in revised form 19 January 2012; accepted 20 January 2012

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 pleasant is a voice from a perceptual point of view when the final application is a speech based interface. In this paper we present an objective definition for voice pleasantness based on the composition of a representative feature subset and a new automatic voice pleasantness classification and intensity estimation system. Our study is based on a database composed by European Portuguese female voices but the methodology can be extended to male voices or to other languages. In the objective performance evaluation the system achieved a 9.1% error rate for voice pleasantness classification and a 15.7% error rate for voice pleasantness intensity estimation.

Keywords: Voice pleasantness; Subtle emotions; Perceptual speech analysis; Text-to-Speech synthesis

1. Introduction

Text-to-Speech (TTS) technology has dramatically improved in the last few years and the inclusion of this mature technology in our daily lives is now a reality. We can easily find examples such as GPSs, smartphones, reading assistants, interactive voice response (IVR) and automotive applications that assist us in several tasks. Intelligibility is fully required for any commercial system and several systems are becoming progressively more natural. However there are users that don't feel fully satisfied with their products and often say that the voice is boring, that the style is not very friendly or even that they dislike the sound of the voice, among others. To try to answer to these complaints, we decided to study the concept of voice pleasantness according to the definitions found in Fellbaum (1998) in which "Pleasantness is the feeling caused by agreeable stimuli", and in Oxford Dictionary (2009) where pleasantness is what

☆ This paper has been recommended for acceptance by 'Björn Schuller, PhD'.

* Corresponding author at: Polytechnic Institute of Porto, Portugal. Tel.: +351 914 216 862; fax: +351 214 218 488.
E-mail address: lcoelho@eu.ipp.pt (L. Pinto-Coelho).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2012.01.006

gives "a sense of happy satisfaction or enjoyment". This is distinct from the concept of attractiveness ("appealing to the senses", "sexually alluring" Oxford Dictionary, 2009) which was also evaluated but is out of the scope of this work.

The concept of voice pleasantness and the suitability of a given voice to be used on a given communicative situation are barely touched by scientific literature (Schröder, 2009). Recently, Campbell (2008) integrated the concept in the area of expressive/affective speech as a form of subtle emotion where it appears as a desired feature in TTS systems while the interaction between speaker and listener is shifted from a typical "read speech towards a more conversational style of speech". To characterize expressive speech we often find references to voice quality and prosody parameters and Montero et al. (1999) and Audibert et al. (2006) showed that these can be combined with distinct weights to convey a given expression. Additionally, voice pleasantness can be seen as a more permanent attribute than the intense and time limited emotions often found in the expressive speech literature. This characteristic allows to envision an integration on a speaker identification framework. From these studies and distinct perspectives we start to see that voice pleasantness is indeed a complex concept that encompasses several dimensions and, in this way, its analysis can benefit from areas such as voice quality, speaker identification and emotion recognition. Based on our theoretical framework we will mainly rely on the later but always with support from the first two.

We have previously explored correlations between subjective ratings and objective features in Braga et al. (2007b) and the creation of the database on which we relied is described in Braga et al. (2007a). A first study on voice pleasantness classification, using a small multilingual corpus, was published in Coelho et al. (2011).

In this work the main goals and novelties are (a) the presentation of an objective definition for voice pleasantness based on the composition of a representative feature subset and (b) the optimization of a voice pleasantness evaluation pipeline that allows to automatically identify occurrences of the concept as well as estimating its intensity.

We have specifically focused on European Portuguese (EP) since we had a very homogeneous database with an extensive base of subjective evaluations. The multi-lingual problem will be analysed in a future work.

1.1. Background

From the listener's point of view, when evaluating the impact of the message as stimuli, there are very few studies but the topic is gaining importance. In first reported studies the authors mainly explore correlations between subjective and objective data and the supporting databases are often small and not specific for the analysis of the voice pleasantness topic. For example, Syrdal et al. (1998) conducted a study in order to check the suitability of a speakers' voice to develop a TTS system, based on the assumption that the perceived quality of a natural voice does not necessarily mean synthesized voice quality. The authors have explored the correlation between acoustic characteristics and the subjective attributes of synthetic speech quality (intelligibility, naturalness and pleasantness). The speakers' selection process is said to have been made empirically although all the candidates (6 females and 9 males) were professional speakers. Another example is the work of Yabuoka et al. (2000) who evaluated psychometric scales and correlated the results with an objective evaluation of the acoustic signal. The analysis was mainly directed to investigate the behaviour of the scales and not to the objective definition of the psychometric scales. The authors showed that the objective variables can be grouped into two factors: one related to "clarity" and the other one related to "sensation". However, concerning the type and structure of the subjective test, very little details were given.

From an emotion perception point of view, Gobl and Chasasaide (2003) and Yanushevskaya et al. (2008) showed that it can be highly influenced by voice quality parameters and Lugger and Yang (2008) have used the same parameters to identify happy, bored, neutral, sad, angry and anxious states with very good results for a set of 10 speakers. However there is still a lack of agreement in existent literature on the definition of the optimal features for characterizing emotions (Schuller, 2010).

On similar fields, we find the work of Biadsy et al. (2008) that explored judgements of charisma in Standard American English and Palestinian Arabic speech in a series of perceptual experiments. The authors investigated acoustic, prosodic, and lexical clues that could identify charismatic speech. From an objective point of view, using acoustic–prosodic features, the authors found, for both cultures, that longer speech segments, with frequent changes in f_0, a dynamic usage of different speaking rates and variations in intensity across intonational phrases, were perceived as being charismatic.

More recently, we can find studies that encompass an extended set of features and rely on machine learning techniques for a deeper analysis. Weiss and Burkhardt (2010) correlated listener's likability ratings with an acoustic analysis using emotionally neutral sentences from emoDB (Burkhardt and Paeschke, 2005). High articulation rate, lower spectral

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

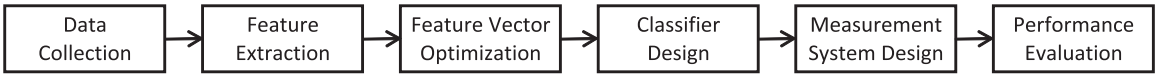


Fig. 1. System's development pipeline.

centre of gravity and higher spectral standard deviation and skewness provided some of the highest correlations for female speakers. In part, these results are aligned with those obtained by Braga et al. (2007a). The authors have also developed a binary classification system that achieved 62.9% accuracy. Burkhardt et al. (2011) have presented results for a binary classification of voice pleasantness using the Agender database (Burkhardt et al., 2010). Despite the large number of speakers, the evaluations were performed using sentences with a command embedded and not always with the same content. The reported accuracy of the system was 67.6%.

1.2. Work overview and document structure

For our purposes we will follow a general machine learning framework. The lack of agreement in existing literature on the definition of the optimal features for characterizing emotions and the scarce number of studies that specifically cover voice pleasantness are the main motivations for the use of this approach since, in the presence of such conditions, it can provide a robust and flexible framework independently of data's nature and allows the exploration of a large number of parameters for system's fine tuning. Our pipeline is organized on a six stage sequence, as depicted in Fig. 1, with function blocks representing the main tasks sets. The specific functions performed on each block will be detailed.

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 and how we analysed the paralinguistic information conveyed in the message in parallel with the psychometric variables to measure the reaction of the listener to the provided stimuli. Afterwards, we will describe the extraction and selection of features as well as the development of a model that could hold the main variables that describe voice pleasantness and how they interact between themselves. Finally, after training and evaluating the system, we will present a thorough discussion of the obtained results and the final conclusions followed by some envisioned future work.

2. Experimental framework

2.1. Database

The database that supported the development of this work is exclusively composed by female voices. These voices belong to professional speakers, with radio, theatre or other vocal experience, and were recorded during voice talent selection processes for the development of new TTS systems. The related recording procedure as well as the quality requirements that were imposed during the processes have been previously published (Braga et al., 2007b) and for this work we will only mention the relevant features. Our database was composed by 77 recordings of distinct speakers, around 3 min of speech each, containing phonetically and prosodically rich sentences which allowed the expression of emotions. The speakers, with EP as mother tongue, had an age range between 20 and 42 years and were all native, speaking the standard dialectal variety (since regional varieties can lead to less positive evaluations Eklund and Lindström, 2001; van Bezooijen, 2005). All the speakers had their own very personal speaking style which allowed to obtain a good diversity for each parameter.

2.2. Subjective evaluation

To evaluate the recorded speech we conducted a survey, using a web application, where each utterance was rated according to pleasantness (and other factors) using a 5 points scale. The exact statement to be rated was "This voice is the most pleasant, has the right melody and is not monotonous at all". The utterances were presented in random order and the listeners were asked to use headphones during the evaluation. The survey received 112 responses, from male and female listeners, native and non-native speakers, in an age range from 23 to 60 years old. The inter-rater agreement

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

was analysed using the Kendal's (Kendall and Smith, 1939) coefficient of concordance. The obtained value was 0.673 which indicates moderate to substantial agreement among listeners.

The subjective evaluation raised several issues related to potential biasing factors. Sex, age, expertise or native/non-native speaker, factors that go beyond the simple selection of parameters, can possibly bias the listeners' judgement analysis. One major concern was the level of expertise on speech processing, since the listeners had distinct backgrounds. A similar problem is addressed by Kreiman et al. (1990) that concludes "that perceptual strategies between more and less experienced listeners are not different, but rather that these listeners adopt different baselines during perceptual tasks". To reduce the group variance the listeners were asked to rate the voices more emotionally rather than using any of their previous experience on the subject.

2.3. Protocol

Using the listeners' voice pleasantness ratings (1–5) we divided the database records into two major classes: pleasant (P) and not pleasant (NP). The first class contained the voices ratted in the two highest scale positions (35 voices) while the second was composed by the voices ranked in the remaining positions (42 voices). The voice pleasantness ratings were used to train the voice pleasantness intensity model while the classes were used to train the classification module. We have used a 10-fold cross validation methodology and on every group we have tried to keep a similar P/NP ratio.

3. Feature extraction and selection

The construction of an optimal feature vector is always a major concern in machine learning problems. On one hand, because it is very difficult, in the presence on a new problem, to identify, from an extremely large set of possibilities, what features will better describe our classes and lead to the best results. On the other hand, because the complexity of most methodologies grows exponentially when the number of features increases, having a reduced feature set can help to significantly reduce the computational requirements.

3.1. Feature extraction

For the study of speech we mainly found two approaches: one, where a small set of descriptors is carefully selected based on the authors expertise and another where a huge number of features is extracted and processed in a posterior stage. In the latest years, with the appearance of several feature extraction tools, such as OpenEAR (Eyben et al., 2009), and with the improvement of feature selection algorithms, we have assisted to an increase on the number of features allowing to explore additional dimensions. For these reasons the second approach is becoming more popular. In our case, since we were working with a small database, we had an extra concern because most machine learning algorithms methods require a sufficient number of records in relation to the number of features in order to provide robust and meaningful results. We decided to use a mixed approach where we first selected a set of features, based on the experience of other authors in this area and in adjacent areas, and second we used a feature selection stage to obtain an optimized subset.

Hence we started with features from the clinical and voice quality areas. The GRBAS (grade, roughness, breathiness, asteny and strain) scale (Pinho and Pontes, 2002) is a typical subjective scale in the clinical field although some objective features are also used. Fundamental frequency and associated dynamics, as well as shimmer, jitter, Harmonic to Noise Ratio (HNR) (Lopes et al., 2008) and Normalized Amplitude Coefficient (NAQ) (Alku et al., 2002), are popular features. From an intelligibility perspective we see that its assessment is highly dependent on the proper articulation of the language sounds according to its standard and that it can vary significantly when dialectal variations are involved, even among speakers of the same language. However Amano-Kusumoto and Hosom (2009) and later Coelho et al. (2010) showed that parameters such as formant trajectories in vowels and segmental durations have paramount importance in speech intelligibility. From the naturalness point of view we have found references to measures of artefacts or discontinuities (Mayo et al., 2011). From the speaker identification point of view we mainly find acoustic features or spectral models which are combined into a speaker model that can be seen as a unique voice signature (Campbell, 1997; Kinnunen and Li, 2010). Finally, from the emotional point of view, there is still a lack of agreement concerning the features that lead to optimal identification results (Schuller, 2010) while the task complexity has enlarged with the high increase on the number of involved features over the years. For example, Ververidis et al. (2004) used 87 acoustic

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

features to recognize 5 emotions, Schuller et al. (2006) used 4000 parameters to identify 7 emotions while in Stuhlsatz et al. (2011) a 6552 dimensional vector was used. Such a large number of features requires very large databases in order to achieve sufficient model robustness and only cross-corpora studies can lead to more general solutions (Schuller et al., 2010; Stuhlsatz et al., 2011).

On the voice pleasantness area there are small differences from the emotion analysis in terms of extracted features. In the work of Syrdal et al. (1998), RMS energy, breathiness, long-term spectra, f_0, formants and their bandwidths, speaking rate, and TTS concatenation and target costs are used for TTS voice pleasantness evaluation. On the other hand, Yabuoka et al. (2000) showed that the objective variables, such as cepstrum distance, amplitude distortion, and waveform distortion, can provide a sensation of "clarity" while phase distortion and differential spectrum distortion can be related to "sensation". Chattopadhyay et al. (2003) used speech rate, pausing and pitch to identify a positive attitude towards voices advertisement. Also in a persuasion related study, Weiss and Burkhardt (2010) used 988 features for correlating voice likability with subjective ratings for classification purposes. Burkhardt et al. (2011) used 4368 features comprising groups of low-level descriptors and a set of functionals for each group. Auditory spectral features provided the best contribution for reliable binary voice likability classification estimates.

Nevertheless and as in Schuller et al. (2007), for most cases a combination of prosody, voice quality and spectral parameters and their dynamics always seem to provide the best results. Hence, 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: (a) acoustic features, (b) signal features, (c) periodicity features and (d) phonation speed features.

For the creation of the prototype feature vector we also took into account the recommendations of Wolf (1972) who advocates that the used features 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 and not be affected by reasonable background noise. In the first group (a), 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, namely /a/, /6/, /i/ and /u/ (using IPA-SAMPA representation) which are the most frequent for EP (Teixeira et al., 2001), we have extracted the first four formants and their related bandwidth ($[V]_i$ represents the frequency of formant i for vowel [V]) and also calculated the six above-mentioned functionals. The second group (b) 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. This information was included in the third group (c) where we have considered jitter (*J*), shimmer (*S*) and harmonic-to-noise ratio (HNR). For each parameter we considered several varieties since they could provide non-redundant information. For jitter we have considered 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). We accounted for six varieties of shimmer: local (Sloc), local in dB (SdB), periodic difference (Sddp) and three amplitude perturbation quotients computed for distinct neighbourhoods using 3, 5 and 11 points. We also used harmonic to noise ratio (HNR) and the same value in dB (HNRdb). All jitter, shimmer and harmonicity parameters were calculated according to Praat's (Boersma, 2001) description. Finally, the last feature group (d) composed of phonation speed metrics, includes the word rate (WR), as words per second, speaking rate (SR), as phonemes per second, 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 parameters we also have considered a speaker model based on 16 MFCCs extracted from 20 ms windows with 5 ms overlaps, considering 4 Gaussian mixtures. This model, based on perceptually weighted parameters, can provide useful information on speaker identification systems that can share several similarities with the voice pleasantness evaluation problem. Since this model can only be considered as a whole we treated it separately and no MFCC features were used during the feature selection stage.

3.2. Multi-criteria feature selection

The prototype feature vector described in the previous section is composed by 179 dimensions. This number brings an increased complexity for the development of the classifier and some of the components may provide little or redundant information for the task. Additionally, from a machine learning point of view, it is necessary to have a

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

Table 1
Composition of prototype feature vector for voice pleasantness analysis.

Group	Feat.	Stat.	#
Acoustic	f_0, Δf_0, $\Delta\Delta f_0$	Av, Std, Kt, Sk, Min, Max	18
	4 × Vow.: 4Fmt	Av, Std, Kt, Sk, Min, Max	64
	4 × Vow.: 4Bw	Av, Std, Kt, Sk, Min, Max	64
Signal	P, ΔP, $\Delta\Delta P$	Av, Std, Kt, Sk, Min, Max	18
Periodicity	Jitter	–	4
	Shimmer	–	6
	HNR, HNRdb	–	2
Phonation speed	WR, SR, PR	–	3

sufficiently large database to ensure an adequate generalization performance of the involved models. This is usually not the case. Since the increase in the number of records may not be easy, it is common to use two distinct categories of algorithms to reduce the size of the feature vector: one operates by transforming the feature space and the other one is a goal oriented search in order to find an optimized feature subset space. In both cases the aim is to maximize the class discriminative power and to reduce information redundancy (Theodoridis and Koutroumbas, 2006). In the first category we can find linear methods, such as *Principal Components Analysis* or *Multidimensional Scaling* (Borg and Groenen, 2005), and non-linear methods, like *Self Organizing Maps* (Kohonen, 2000) or *ISOMap* (Tenenbaum et al., 2000). van der Maaten and Hinton (2008) and van der Maaten and van den Herik (2009) have compared 32 popular dimensionality reduction techniques and have concluded that "nonlinear techniques are often not capable of outperforming traditional techniques such as PCA". In the second category there is a wide variety of methodologies and the large amount of published articles, often contradictory, makes it difficult to clearly select the option that will best suit a given problem or dataset. Zhao et al. (2009) performed a comparison of twelve commonly used feature selection methodologies using several distinct datasets and showed that simple methodologies, such as the *t*-test, can perform as good as other modern and more complex approaches, like the *Kruskal–Wallis* methodology (Cover and Thomas, 1991). In the emotion recognition arena, Luengo (2010) has developed a system centred in the basic Ekman's big six (Ekman et al., 1969), that achieved good results using the *mRMR* feature selection methodology (Peng et al., 2005) (also included in the study performed by Zhao et al. (2009)). The use of Sequential Forward Selection (Ververidis et al., 2004) and related algorithms, such as Sequential Forward Floating Selection (Pudil et al., 1994; Hassan and Damper, 2009; Lugger and Yang, 2008), is also popular but it can lead to sub-optimal solutions.

In our case we wanted to obtain an optimal subset that could only be guaranteed by an exhaustive search. However, with such a large number of features, the direct application of this approach would be computationally unfeasible. This way we had to apply a multi-criteria methodology for feature pre-processing that we designated by multi-criteria feature selection (MCFS). Our pipeline was composed by four stages: feature normalization, feature pre-selection, composite feature selection and exhaustive search selection. Hence, considering that the values on each dimension followed a Gaussian distribution, we 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 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 (Lilliefors, 1967) 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 started the composite feature selection by first creating a feature list sorted according with the feature's class-discriminatory power. Adding into account the inter-feature correlation we have iteratively created a new list where each feature was selected by having the highest class-discriminatory and the lowest correlation with the previously selected features. From this list we have selected the 12 highest ranked features and performed an exhaustive search for combinations of n features using scatter matrices (Fukunaga, 1990). As cost function for class separability measurement we used the J_3 criteria defined as:

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

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

where $\mathbf{S}_w$ is the intraclass scatter matrix and $\mathbf{S}_b$ is the interclass scatter matrix. For each n we retained the highest scored subset. This process allows obtaining a feature list where the best ranked features maximize the discriminative power and minimize the redundancy. To evaluate the performance of our approach we compared the results with the ones obtained using PCA and ISOMap, two popular methodologies.

4. Classification and intensity estimation

Voice pleasantness classification is a specific pattern recognition problem under the machine learning topic. In this case, the general task of assigning a label or value to a given input pattern appears in the form of defining if a voice is pleasant or not using a set of objective features extracted from a given input audio stream. The voice pleasantness intensity measurement task involves measuring the degree of pleasantness when a voice is classified as being pleasant or not pleasant.

4.1. Classification

From the emotion recognition area, that has similarities with our problem, we often find, for classification purposes, methodologies based on Artificial Neural Networks (ANN) (Unluturk et al., 2009) or Support Vector Machines (SVM) (Shaukat and Chen, 2008), with the later being more popular. In fact, Shami and Verhelst (2007) and Casale et al. (2008) performed a comparison of several classification techniques and obtained the best results using SVMs. More recently, new methodologies, such as LogitBoost Alternating Decision Trees (Weiss and Burkhardt, 2010) or ensembles of REPTrees (Burkhardt et al., 2011), released as plug-ins for the WEKA toolkit (Witten and Frank, 2005), are also becoming popular. However the performance of the new approaches may be highly conditioned by the data distribution while ANN and SVM often offer a more consistent good performance across databases (Schuller et al., 2010). On the other hand, Hassan and Damper (2009) reported to beat the best reported classification results on the Berlin and DES emotion databases using a strategy based on a simple classifier (k-nearest neighbour). In our case we considered the ANN and SVM approaches, due to their robustness and flexible models, and compared their performance.

Since voice pleasantness can be a more permanent characteristic and, assuming that it is not manipulated or affected by an ill state, it can also be understood as being intrinsic to the speaker. In this sense we also evaluated the suitability of a speaker identification model where the speaker's acoustic signature is used to represent voice pleasantness identity. For this we used a Bayesian classifier and a speaker model based on Gaussian mixtures (GMM), which can lead to good results (Beigi, 2011). The implementation was inspired on Reynolds et al. (2000).

The ANN and GMM based approaches were implemented using Matlab scripts. The ANN had a feed-forward structure with 2 hidden layers and 14 neurons each (this configuration was obtained by optimizing the cost function). The size of the input layer was equal to the dimension of the input vector (which was one of the varied parameters). For output we had a single neuron, with a sigmoid activation function, where 0 and 1 represented not pleasant and pleasant, respectively. The back-propagation algorithm was used for training. The GMM model was built as described in the features section and the Expectation Maximization (EM) (Dempster et al., 1977) algorithm was used for training. The SVM approach was based on the implementation provided by the LibSVM (Chang and Lin, 2011) toolkit using a C-SVC SVM model and radial basis functions with a gamma value equal to 8 and a cost parameter of 2 (these parameters were adjusted by minimizing classification error with a grid search methodology).

The classifiers' training/testing processes was performed using a 10-fold cross validation that allowed to obtain statistically meaningful results. A late fusion based on a weighted voting scheme, adjusted by calculating the relative number of true positives for each classifier against the total number of true positives, allowed to reach a hybrid model for category estimation.

4.2. Voice pleasantness intensity estimation

Besides identifying voice pleasantness it was also our goal to measure its intensity. This is different from the concept of "ranking", that often appears in the machine learning literature, where the probability associated with the binary classification decision is used to provide additional information concerning its trustfulness. This can be viewed as an "intensity estimation" for the class but for our case this is of limited use. For example, when a given point is located very far from the decision boundary and right inside the class, it will have associated a high probability but it might

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

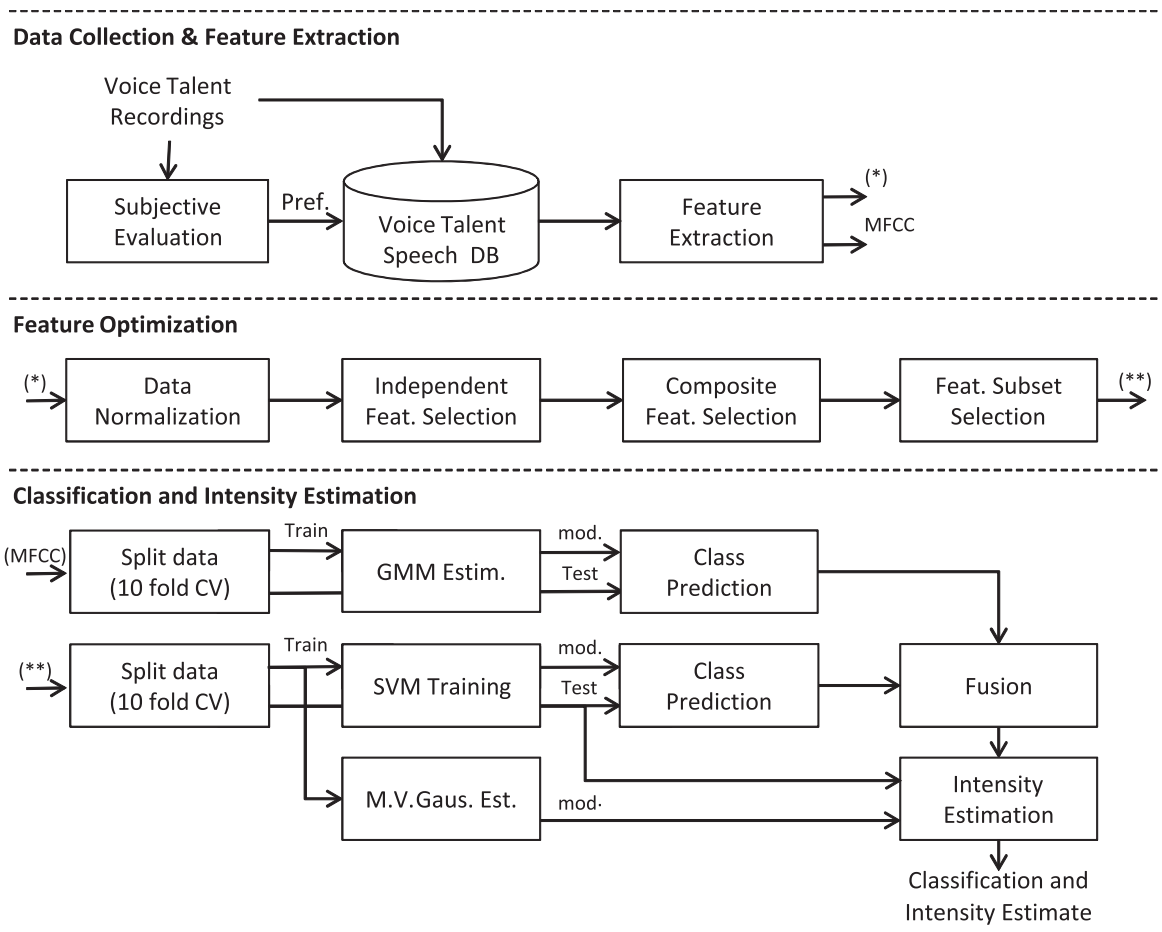


Fig. 2. System architecture.

not represent an optimal location in terms of voice pleasantness. The feature selection process tested features for normality thus making the data Gaussian compatible. Hence, for voice pleasantness intensity estimation we have used a multivariate Gaussian distribution whose parameters were trained with voice judgement data from the two classes and using the optimized feature vector previously devised. Each class has its own model that is chosen after having the voice pleasantness class estimation. This approach allowed to use a nonlinear class discrimination scheme that will first divide the feature space and then will use a class dependent model, based on a multivariate Gaussian distribution, to reach a final voice pleasantness intensity estimation.

As an alternative scenario we have used only the first two best features/components calculated in accordance with the presented methods. For each parameter pair (p_1, p_2) we associated a pleasantness ranking as a dependent function $r=f(p_1, p_2)$. Then using a uniform bicubic B-spline surface $S(x, y)$, we started with a plane that was successively deformed in order to approximate the true ranking value for the given points (Barsky, 1982). For each class we have a distinct surface that was adjusted to the related points. This approach can provide good results though it poses some limitations due to its non-parametric nature.

5. Evaluation and results

The best results were obtained using the configuration depicted in Fig. 2. This optimized pipeline allowed to achieve a final voice classification error rate of 9.1 ± 2.1 % and a voice pleasantness intensity estimation absolute error rate of 15.7 ± 2.5, both for a 90% confidence interval.

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

Table 2
List of the first 84 features in the ranked list.

	0	12	24	36	48	60	72	84
1	**Jrap**	f_0Min	VoB4Sk	VeF1Sk	VaB1Kt	VeF1Kt	NHR	ViF2Max
2	SR	$\Delta\Delta P$Sk	VeB1Av	f_0Max	VeB3Max	VeB3Min	ViF1Max	VaF3Max
3	VaB1Av	VaF2Kt	VaF1Std	VoB4Av	VaB1Max	VoB4Max	VoF2Kt	VaB3Kt
4	**ΔPAv**	VoB3Std	VeF3Max	VoF3Min	VoB4Std	ViF3Std	ViB2Max	ViF2Sk
5	**Δf_0Sk**	ΔPMin	VoB3Av	VoB1Av	VaB2Std	VeF4Min	ViF2Min	VeB1Sk
6	f_0SD	VoB3Sk	ΔPSD	$\Delta\Delta f_0$SD	ViF4Min	VeB4Min	VoF4Sk	VeF3Min
7	**Sapq3**	$\Delta\Delta P$Max	PR	VoB1Min	VeF4Sk	ViF3Min	VoF2Sk	Sdda
8	HNR (dB)	VaB4Min	HAC	VaB2Max	ViF1Av	VeF2Std	VeB3Sk	Jddp
9	**f_0Av**	VeB4Kt	f_0K	VaF4Kt	ViB2Av	VoB3Min	VeF3Av	VoF4Max
10	Δf_0Av	JLAbs	VeB2Kt	$\Delta\Delta P$Av	PAv	Δf_0Min	VoF4Min	VaB3Max
11	PKt	VeB3Std	JLAbs2	VoB2Sk	ViF2Kt	VoF1Std	VeF2Sk	VaB1Sk
12	**Δf_0Max**	ΔPSK	VoF2Std	VaF4Sk	ViF1Min	VoB1Std	VaB4Std	VoB3Max

Table 3
Reference values for the best ranked features. Values are absolute with a relative variation within a 90% confidence interval.

Feature	Jrap	ΔPAv	Δf_0Sk	Sapq3	f_0Av	Δf_0Max
Value	$0.00946 \pm 9.3\%$	$0.0110 \pm 12.1\%$	$0.576 \pm 4.7\%$	$0.02697 \pm 5.9\%$	$196 \pm 5.3\%$	$34.24 \pm 6.2\%$

The system optimization was performed by individually adjusting each of the parameters on every stage while observing the performance improvements in the output. The performance was always measured in terms of classification error (sum of false positives and false negatives over the total number of evaluated cases) and a 10-fold cross validation allowed to reach the mean error rate as well as the confidence interval.

We will follow the pipeline from the beginning. After extracting the features, we have applied the proposed MCFS methodology. The ranked feature list, after composite feature selection, is presented in Table 2.

From the ranked feature list we went to investigate the amount of necessary features to provide the best results. Using a SVM classifier with a RBF kernel, whose gamma and cost parameter were optimized in each experiment, we tested several combinations of n out of the 12 best ranked features in the list. The best results were obtained using a 6 elements vector composed by the features represented in bold face in the first column of Table 2. The related values are presented in Table 3. Additionally we wanted to evaluate if it was possible to obtain better results using dimensionality reduction approaches such as PCA or ISOMap. In Fig. 3 we can observe the obtained classification error and the related 90% confidence interval for the three approaches while varying the number of features or vector components. We can see that PCA and ISOMap clearly outperform MCFS when a reduced number of components are considered and that the evaluation of a small set of independent features cannot provide a sufficient discriminant power. The performance of all methods improves with the size of the feature vector until a certain point with PCA and ISOMap achieving the best results with 5 components and MCFS with 6 features and the overall smallest error rate. When considering a larger feature vector we can see that the MCFS performance is degraded which may indicate difficulties during classifier

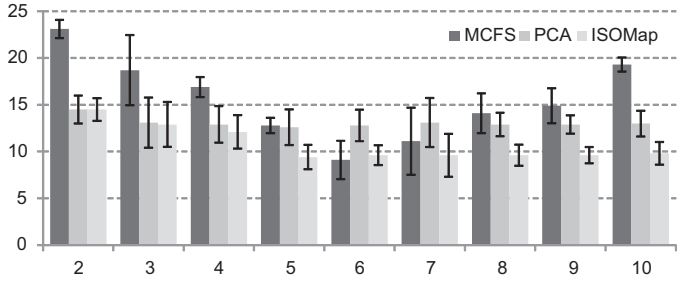


Fig. 3. Classification error according with feature vector optimization methodology and number of extracted features, in percentage.

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

Table 4
Composition of the first PCA component.

Component	1	2	3	4	5	6
Feature	SR	Sapq3	Δf_0Av	ΔEMax	Jrap	VeB3Kt
Weight	12.2	10.8	8.5	5.5	5.3	3.9
Acc. weight	12.2	23.0	31.5	37.0	42.3	46.2

Component	7	8	9	10	11	12
Feature	f_0SD	$\Delta\Delta E$SK	VeB1Av	HAC	$\Delta\Delta E$Max	PR
Weight	3.7	3.0	2.8	2.7	2.5	2.3
Acc. weight	49.9	52.9	55.7	58.4	60.9	63.2

training when dealing with additional dimensions. On the other hand, PCA and ISOMap, showed to be much more robust to vector size.

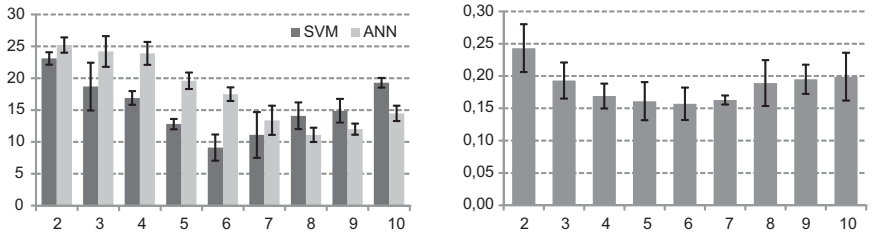
Since PCA and ISOMap performed well with small sized description vectors and we wanted to understand the features that best represented voice pleasantness we investigated the composition of the first PCA component. In Table 4 we see the 12 highest weighted features as well as their accumulated weight. We can observe that most of these features are also among the highest ranked in the MCFS feature list (Table 2) which reinforces their importance on the definition of voice pleasantness. It is also noticeable that the most important features in both approaches are mainly voice quality and prosody parameters. The average f_0, the maximum Δf_0, speaking rate and pause rate, among others, are also listed as having a high correlation with voice likability in Braga et al. (2007a), which can bring additional confidence in the results.

In the classification stage, and still thinking in the previous stage, we compared the performance of the SVM and ANN approaches while varying the feature vector dimension. The results can be seen in Fig. 4(a) where we observe that the SVM classifier still holds the best performance. (The GMM classifier was evaluated separately and we have not varied the dimension of the input vector.)

For classification we evaluated independently three classifiers, ANN, SVM and GMM as can be seen on Table 5(a)–(c), respectively. The SVM classifier performed better and to optimally tune the SVM classifier, we evaluated the system's performance when varying the type of kernel function and related parameters, as shown in the first row of Table 6 (before fusion). 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).

In Table 5(c) we can observe that the GMM–Bayes classifier showed to have a better performance when identifying the not so pleasant voices. To further improve the system we combined the SVM classifier, which achieved the best overall performance, with the GMM/Bayes classifier using a weighted voting scheme (performed during training). This hybrid approach allowed to improve the overall results, which are shown in the second row of Table 6. For the RBF kernel, a relative weight of 30% GMM and 70% SVM was obtained by an error minimization search procedure.

Additionally we also evaluated in what extent the feature selection process and the inclusion of a second parallel classifier (GMM) helped to improve the classification results. In Table 7, we can observe the improvements introduced



(a) Classification error (b) Voice pleasantness intensity estimation error

Fig. 4. Performance evaluation according to feature vector dimension. Horizontal axis indicates the dimension of the feature vector and vertical axis shows error in percentage (a) and absolute value (b).

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

Table 5
Confusion matrices with classification results, in percentage, for distinct approaches. Classes are pleasant (P) or not pleasant (NP). On the left we have the real classes (R) and on the first row we have the estimated classes (E).

R\E	P	NP
(a) ANN		
P	80.5	21.1
NP	19.5	78.9
(b) SVM		
P	89.8	12.7
NP	10.2	87.3
(c) GMM		
P	68.5	11.4
NP	31.5	88.6

Table 6
Voice preference classification error for a 90% confidence interval using different kernel types with a SVM based classifier and after fusion with a GMM/Bayes classifier. (Two classes outlier free dataset, error within a 90% confidence interval.)

	Linear	Quadratic	RBF	Sigmoid
SVM alone	23.7 ± 3.7%	18.8 ± 3.2%	11.5 ± 2.6%	24.2 ± 3.9%
Class. fusion	19.3 ± 3.7%	14.5 ± 2.1%	9.1 ± 2.1%	21.9 ± 4.2%

Table 7
Classification error rate after the introduction of each component. (Error within a 90% confidence interval.)

Non-tuned system	Independent feat. sel.	Composite feat. sel.	Optimized feat. subset	Composite classification
38.5 ± 7.6%	18.9 ± 5.2%	14.4 ± 2.3%	12.3 ± 3.1	9.1 ± 2.1%

by each block of the pipeline using as reference the results obtained with a RBF kernel (since it performed better) and considering all the available objective features.

After obtaining a classification result we performed the voice pleasantness intensity estimation. The performance evaluation was based on the mean absolute estimation error, or by other words, the mean absolute difference between the true human judged pleasantness rank (on a 1–5 scale) and the value estimated by the system. The results obtained for the multivariate Gaussian model approach are depicted in Fig. 4(b) where we can observe that the best results, 15.7± 2.5 % for a 90% confidence interval, are obtained again for a 6 dimensions vector. The surface based approach was implemented using a 20 × 20 patch grid supported on the best two ranked features using the MCFS process and the first two components/vectors of the PCA and ISOMap methodologies. The obtained error rates were 27.3± 5.7 %, 15.8± 3.2 % and 15.6± 3.9 %, respectively, for a 90% confidence interval. A surface mapped from a bidimensional reference based on raw features was clearly inappropriate however, the use of PCA or ISOMap, led to successful results. Especially the ISOMap based reference provided a residually lower error rate but with a higher variation in the confidence interval. For this reason we preferred the parametric approach with better repeatability.

6. Conclusions and future work

In this paper we have presented an optimized pipeline for the automatic evaluation of voice pleasantness. We started by defining the pleasantness concept, which has a central role in our work, and showed how it can be related to the perception of voice as stimuli. An extensive presentation of the most significant studies in the area was also presented. After showing an overview of the processing pipeline, we have detailed the functions of each stage. We provided a brief presentation of the database, which was specifically developed for the study of voice pleasantness. Then, we thoroughly explained the procedure that was followed to obtain an optimal feature vector for achieving improved results. We started by exploring a broad range of dimensions, covering acoustic, signal, periodicity and phonation speed aspects

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

and successively selected the best features by maximizing their class discriminatory power and by reducing the inter-feature redundancy. This approach provided the best results for our data after a comparison with other methodologies. Our optimized feature vector, mainly composed by prosodic and voice quality parameters, included **Jrap** (relative average perturbation), **Δ*P*Av** (derivative of average power), **Δf_0Sk** (skewness of fundamental frequency derivative), **Sapq3** (amplitude perturbation quotient), **f_0Av** and **Δf_0Max**. A SVM classifier was successfully combined with a GMM/Bayes methodology in a late fusion scheme. This new approach, motivated by the specific nature of voice pleasantness analysis, allowed to achieve a final voice pleasantness classification error rate of 9.1 ± 2.1 %, which is better than other reported marks for similar tasks. Finally, for estimating voice pleasantness intensity, we achieved the best results using a class dependent multivariate Gaussian model, which provided an error rate of 15.7 ± 2.5 %. The obtained results were obtained using a database that was specifically developed for the study of voice pleasantness.

The presented system is a powerful objective analysis tool that can support subjective evaluations during voice talent selection stage for new voice fonts recording, thus contributing to reduce overall TTS development costs. This work also provides an additional insight on the study of pleasantness, one of the key dimensions of expressive speech, and can contribute to the enhancement of TTS systems' voice quality as well as to leverage the introduction of this technology in more real life applications.

As future work we plan to apply a similar methodology to other languages allowing to simultaneously increase the available data and to evaluate the cross-language performance. An extension of the work for characterizing male voices is also envisioned.

Acknowledgements

This work was partially supported by ERDF funds, the Spanish Government (TEC2009-14094-C04-04), and Xunta de Galicia (CN2011/019, 2009/062).

References

Alku, P., Backstrom, T., Vikman, E., 2002. Normalized amplitude quotient for parametrization of the glottal flow. Acoustical Society of America 2, 701–710.

Amano-Kusumoto, A., Hosom, J.-P., 2009. The effect of formant trajectories and phoneme durations on vowel intelligibility. In: Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing, pp. 4677–4680.

Audibert, N., Vincent, D., Aubergé, V., Rosec, O., 2006. Expressive speech synthesis: evaluation of a voice quality centered coder on the different acoustic dimensions. In: Proc. of 3rd International Conference on Speech Prosody, Dresden, Germany.

Barsky, B.A., 1982. End conditions and boundary conditions for uniform b-spline curve and surface representations. Computers in Industry 3, 17–29.

Beigi, H., 2011. Fundamentals of Speaker Recognition. Springer.

Biadsy, F., Rosenberg, A., Carlson, R., Hirschberg, J., Strangert, E., 2008. A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech. In: Proc. Speech Prosody, Campinas, Brazil.

Boersma, P., 2001. Praat: doing phonetics by computer. Glot International 9/10 (5), 341–345, URL http://www.fon.hum.uva.nl/praat/.

Borg, I., Groenen, P., 2005. Modern Multidimensional Scaling: Theory and Applications. Springer-Verlag.

Braga, D., Coelho, L., Junior, F.G.R., Dias, M.S., 2007, July. Subjective and objective evaluation of Brazilian Portuguese TTS voice font quality. In: Int. Workshop on Advances in Speech Technology, Maribor, Slovenia, pp. 306–311.

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

Burkhardt, F., Eckert, M., Johannsen, W., Stegmann, J., 2010. A database of age and gender annotated telephone speech. In: Proc. Languages Resources Evaluation Conference.

Burkhardt, F., Paeschke, A., 2005. A database of German emotional speech. In: Proc. Interspeech.

Burkhardt, F., Schuller, B., Weiss, B., Weninger, F., 2011, August. "Would you buy a car from me?" – on the likability of telephone voices. In: Proc. Interspeech, Florence.

Campbell, J.P., 1997. Speaker recognition: a tutorial. Proceedings of the IEEE 85 (9), 1437–1462.

Campbell, N., 2008. Expressive/affective speech synthesis. In: Springer Handbook on Speech Processing. Springer, http://www.speech-data.jp/fastnet/ApplicantCV-Campbell.pdf, pp. 505–517.

Casale, S., Russo, A., Scebba, G., Serrano, S.,2008. Speech emotion classification using machine learning algorithms. In: Proceedings of the 2008 IEEE International Conference on Semantic Computing. IEEE Computer Society, Washington, DC, USA, pp. 158–165.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27.

Chattopadhyay, A., Dahl, D.W., Ritchie, R.J., Shahin, K.N., 2003. Hearing voices: the impact of announcer speech characteristics on consumer response to broadcast advertising. Journal of Consumer Psychology 13 (3), 198–204.

Coelho, L., Braga, D., Dias, M., Mateo, C., 2011. An automatic voice pleasantness classification system based on prosodic and acoustic patterns of voice preference. In: Proc. of Interspeech.

Coelho, L., Braga, D., Garcia-Mateo, C., 2010. Kalman tracking linear predictor for vowel intelligibility enhancement on European Portuguese HMM based speech synthesis. In: Proc of ICASSP. No. 1, pp. 4734–4737.

Cover, T., Thomas, J., 1991. Elements of Information Theory. John Wiley & Sons, New York.

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

Eklund, R., Lindström, A., 2001. Xenophones: an investigation of phone set expansion in Swedish and implications for speech recognition and speech synthesis. Speech Communication 35 (2), 81–102.

Ekman, P., Sorenson, E., Friesen, W.V., 1969. Pan-cultural elements in facial displays of emotions. Science 164, 86–88.

Eyben, F., Wöllmer, M., Schuller, B., 2009. OpenEAR – introducing the Munich open-source emotion and affect recognition toolkit. In: Proc. 4th Int. HUMAINE Association Conf. on Affective Computing and Intelligent Interaction.

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

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press.

Gobl, C., Chasasaide, A.N., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40, 189–212.

Hassan, A., Damper, R.I., 2009. Emotion recognition from speech using extended feature selection and a simple classifier. In: Proc. Interspeech, Brighton.

Kendall, M.G., Smith, B., 1939. The problem of m rankings. The Annals of Mathematical Statistics 10 (3), 275–287.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Communication 52 (January), 12–40.

Kohonen, T., 2000. Self Organizing Maps. Springer.

Kreiman, J., Gerratt, B., Precoda, K., 1990. Listener experience and perception of voice quality. Journal of Speech and Hearing Research 33, 103–115.

Lilliefors, H.W., 1967. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62, 399–402.

Lopes, J., Freitas, S., Sousa, R., Matos, J., Abreu, F., Ferreira, A., 2008. A medida HNR: Sua relevância na análise acústica da voz e sua estimação precisa. In: Proceedings of "I Jornadas sobre Tecnologia e Saúde", Guarda, Portugal.

Luengo, I., 2010. Análisis y evaluación de parámetros para identificación automática de emociones en el habla. Ph.D. Thesis, Universidad del País Vasco.

Lugger, M., Yang, B., 2008. cal motivated multi-stage emotion classification exploiting voice quality features. In: Speech Recognition, Technologies and Applications, pp. 395–410.

Mayo, C., Clark, R.A.J., King, S., 2011. Listeners' weighting of acoustic cues to synthetic speech naturalness: a multidimensional scaling analysis. Speech Communication 53 (March), 311–326.

Montero, J.M., Gutiérrez-Arriola, J., Colas, J., Enriquez, E., Pardo, J.M., 1999. Analysis and modelling of emotional speech in Spanish. ICPhS99 2, 957–960.

Oxford Dictionary, 2009 [Online; accessed 13.03.2011]. URL http://oxforddictionaries.com/.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1226–1238.

Pinho, S., Pontes, P., 2002. Escala de avaliação perceptiva da fonte glótica: Rasat. Vox Brasilis 8 (3), 8–13.

Pudil, P., Ferri, F., Novovicova, J., Kittler, J., 1994. Floating search method for feature selection with nonmonotonic criterion functions. Pattern Recognition 2, 279–283.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41.

Schuller, B., 2010. On the acoustics of emotion in speech: desperately seeking a standard. Journal of the Acoustical Society of America 127 (3), 1995.

Schuller, B., Arsic, D., Wallhoff, F., Rigoll, G., 2006. Emotion recognition in the noise applying large acoustic feature sets. In: Proc. Speech Prosody, Dresden.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kossous, L., Aharonson, V., 2007, August. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Proceedings of Interspeech, Antwerp, Belgium.

Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G., 2010. Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Transactions on Affective Computing 1 (July), 119–131.

Schröder, M., 2009. Expressive speech synthesis: past, present, and possible futures. In: Affective Information Processing. Springer, London, pp. 111–126.

Shami, M., Verhelst, W., 2007. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Communication 49 (March), 201–212.

Shaukat, A., Chen, K., 2008. Towards automatic emotional state categorization from speech signals. In: INTERSPEECH'08, pp. 2771–2774.

Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, H.-G., Schuller, B.,2011. Deep neural networks for acoustic emotion recognition: raising the benchmarks. In: Proc of ICASSP. IEEE, pp. 5688–5691.

Syrdal, A., Conkie, A., Stylianou, Y., 1998. Exploration of acoustic correlates in speaker selection for concatenative synthesis. In: Proc. of Int. Conf. Speech and Lang. Processing 98, Lisbon, Portugal.

Teixeira, J.P., Freitas, D., Braga, D., Barros, M.J., Latsch, V., 2001. Phonetic events from the labelling the European Portuguese database for speech synthesis, FEUP/IPB-DB. In: Proc. of Eurospeech.

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323.
Theodoridis, S., Koutroumbas, K., 2006, February. Pattern Recognition, 3rd ed. Academic Press.
Unluturk, M.S., Oguz, K., Atay, C.,2009. Emotion recognition using neural networks. In: Proceedings of the 10th WSEAS International Conference on Neural Networks. WSEAS, Stevens Point, WI, USA, pp. 82–85.
van Bezooijen, R., 2005. Approximant /r/ in Dutch: routes and feelings. Speech Communication 47 (1), 15–31.
L.J.P. van der Maaten, E. P., van den Herik, H., 2009. Dimensionality reduction: a comparative review. Tech. rep., Tilburg University Technical Report.
van der Maaten, L., Hinton, G., 2008. Visualizing high-dimensional data using *t*-sne. Journal of Machine Learning Research 9, 2579–2605.
Ververidis, D., Kotropoulos, C., Pitas, I., 2004. Automatic emotional speech classification. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), vol. 1, pp. 593–596.
Weiss, B., Burkhardt, F., 2010, September. Voice attributes affecting likability perception. In: Proc. of Interspeech, Makuhari, Japan.
Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
Wolf, J.J., 1972. Efficient acoustic parameters for speaker recognition. Journal of the American Statistical Association 51, 2044–2056.
Yabuoka, H., Nakayama, T., Kitabayashi, Y., Asakawa, Y., 2000. Investigations of independence of distortion scales in objective evaluation of synthesized speech quality. Electronics and Communications in Japan (Part III: Fundamental Electronic Science) 83 (5), 14–22.
Yanushevskaya, I., Gobl, C., Chasaide, A.N., 2008. Voice quality and loudness in affect perception. In: Proc. of Speech Prosody.
Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., Liu, H., 2009. Advancing feature selection research – ASU feature selection repository. Tech. rep.

Please cite this article in press as: Pinto-Coelho, L., et al., On the development of an automatic voice pleasantness classification and intensity estimation system. Comput. Speech Lang. (2012), doi:10.1016/j.csl.2012.01.006