Towards an Objective Voice Preference Definition for the Portuguese Language

Luis Coelho [1], *Horst-Udo Hain*[2], *Oliver Jokisch*[2] *and Daniela Braga*[3]

[1]ESEIG, Instituto Politecnico do Porto, Porto, Portugal
[2]Laboratory of Acoustics and Speech Communication, TU Dresden, Germany
[3]Microsoft Language Development Center, Microsoft, Portugal

lcoelho@eu.ipp.pt, [horst-udo.hain, oliver.jokisch]@tu-dresden.de, i-dbraga@microsoft.com

Abstract

In this paper, it is our aim to define a set of objective acoustic criteria, based on subjective listeners' assessment of talent voices, which can help to automatically rate the voice font quality, bearing in mind the objective definition of voice preference for the Portuguese language. For this purpose a multilingual and multispeaker database was recorded and a set of subjective and objective information was obtained. The analysis of the data provided new results that can be successfully used to define the quality of a given voice. The achieved results for Portuguese were compared with those obtained for other language with objective of identifying common properties, which was statistically confirmed with a within a 90% confidence interval.

Index Terms: voice pleasantness, speech synthesis

1. Introduction

The possibility to create an artificial voice that could imitate a human speaking is slowly becoming a reality. The developments in the last few years made it possible for Text to Speech (TTS) Systems to generate highly intelligible speech with an almost natural prosody. The industry started to take advantage of this ready to use technology and several systems started to emerge from the laboratories to personal computers, cars, and more recently to several applications on mobile devices. However with the fulfilment of the basic requirement, which is intelligibility, additional demands arise. For a daily usage of such technology the system must be robust, to transmit confidence, and the quality of the used voice font must be sufficient to provide a pleasant experience during interaction. Several companies provide for each language several voice fonts that the user can choose according to his preference, but usually this is not the case.

There are few known studies concerning voice quality assessment according to voice preference. This concept is often associated in specialized literature with impairments or disorders and is mostly covered on medical publications. For this subject there is an extensive bibliography, but voice quality is understood, on a positive way, as a voice with no associated pathologies. The evaluation of such voices is performed by using specific metrics that measure the deviation in relation to a range of pre-defined values that define a healthy condition. Standard scales as the GRBAS (grade, roughness, breathiness, asteny and strain) or RASAT (roughness, asperity, breathiness, asteny and tension) [1] are used to help experienced professionals to subjectively evaluate and classify the severeness of a voice dysfunction. Characteristics as hoarseness, raspiness, effort to talk, vocal fry, uncomfortable or abnormal pitch and other abnormal vocal symptoms are also commonly evaluated [2, 3, 4]. These parameters are not automatically extracted, it is required a human subjective evaluation whose judgement can be controversial. As reported by other authors the subjective judgment of distinct professionals does not always present an expressive correlation [5, 6] and there are no guidelines or references for performing the evaluation.

In this paper we explore the voice quality concept on the dimension of voice preference specifically for the Portuguese language, European (EP) and Brazilian (BP) varieties. For this purpose we collected an extensive voice database with professional voices essentially from the media industry. Among several rules, the voice selection was performed to eliminate other factors than the acoustics itself. It is known that accent and even dialect, new words, maybe even phraseology changes, such as between UK and United States, can lead to undesired bias on voice evaluation. The recordings were evaluated by groups of listeners according to voice preference. To find objective voice preference clues we also extracted acoustic parameters and correlated obtained values with the subjective voice rankings. Additionally in a cross-lingual study we further extended the initial recordings to other languages and performed new evaluation surveys. Unlike voice talents, the human evaluator were selected in order to create a heterogeneous group according to external parameters such as age and gender. These variations provided indirect analysis that enriched the results. The findings for EP and BP were compared with the ones obtained for the other languages and statistically significance tests were performed.

The rest of the paper is organized as follows. In the next section we briefly describe speaker selection and database recording processes. The used criteria for voice talent selection and the specific related issues are presented along with the used recording structure. In section 3 we show how we proceeded to evaluate the voices by describing the process as well as the subjective and objective parameters used for this purpose. In section 4 we present the main outcomes for the independent analysis and for the comparison between EP+BP and the other languages. Finally, in section 5 the main conclusion are presented and envisioned work is foreseen.

2. Speech Resources

Our initial studies are based on two voice talent selection processes for European and Brazilian Portuguese with the aim of building a high quality voice font [7, 8] for a new TTS system. During the voice assessment process, which will be explained in the next section, the candidates were asked to record a small text on a professional recording studio, to guarantee identical acoustical quality, while following a common script containing a set of phonetically and prosodically rich sentences, with emotion indications. The voice assessment process followed a well defined pipeline with strict rules and organized in three stages. The first stage was a national call for voice talents which had to fulfil a few profile requirements. Each candidate had to be a female, have Portuguese as mother tongue, having studied up to university level, speaking accent according to the national standard and to have some radio or theatre vocal experience. Out of several hundred candidates, a small set was invited to send samples of their voices with the maximum quality they could produce. A subjective test was then conducted, using a 5 points rating MOS scale, with listeners who were familiarized with speech processing technology. The best scored candidates were then invited to record a small text as described in section 2. The final recordings were evaluated again by a survey where then listener elected the best voice for each attribute. The final ranking was obtained by counting the number of votes each voice received during the survey (further details can be found on [7]). For extending the study base, similar procedures were conducted for Catalan (ES-CAT), Danish (DAN) and Finnish (FI) which allowed to establish comparisons and improve the confidence on the results.

A set of recordings performed within a cooperation between Siemens AG, Munich, and TU Dresden for the creation of new voices for an embedded version of the multi-lingual TTS system ”Papageno” was also used [9]. Amongst others, voices for German (GE), UK and US English (ENG-UK and ENG-US), French (FR) and Spanish (ES) have been recorded at TU Dresden laboratories. As before, all the speakers were selected ensuring that a set of requirements was fulfilled. In general, the voice has to be intelligible, natural and pleasant. A special demand is that it must be suitable for all the processing steps that are involved in speech synthesis. The voice quality (F0, jitter) must be sufficient and allow for good results even after compression or codecs (e. g. adaptive multi-rate, AMR) are applied. The speaker needs to have phonetic and also prosodic abilities (preferable a professional or semi-professional speaker) and should have experience in speaking a long time (about 4 hours per session) without any degradation of the voice quality (e. g. a teacher, actor, newsreader).

All the acoustic data was recorded in professional studios with a sampling rate of 44.1kHz or higher (mono channel) and with 16 bits resolution. From all the recordings a set of randomly chosen sentences was selected in order to obtain around 5 minutes of speech per speaker. For each language at least 5 speakers were considered and for EP and BP the sample was constituted by 10 speakers each.

The described recordings and related scripts, despite their distinct origin, were made using identical criteria. The data organization enabled us to create a homogeneous basis for our analysis.

3. Evaluation

The evaluation of each voice was performed according to subjective and objective parameters, correlated afterwards.

3.1. Subjective Parameters

The subjective evaluation raised several issues related with potential biasing factors. Sex, age, expertise or native/non-native speaker, factors that go beyond the simple selection of parameters, can dramatically bias the listeners' judgment analysis. One major concern was the level of expertise on speech processing knowledge since the listeners had distinct backgrounds. This problem is addressed in [10] and more recently in [6]. In the former ratings from speech and language therapists specialized in voice with at least 2 years experience are compared with those of final year speech and language therapy students. In total 14 parameters like breathiness, roughness and monotony as well as pitch or loudness were investigated. An important basic condition is that only the perceptual labels that are reliably judged by both listener groups should be used for comparison. The author 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 it was asked to the listeners to rate the voices more emotionally rather than using any of their previous experience on the subject.

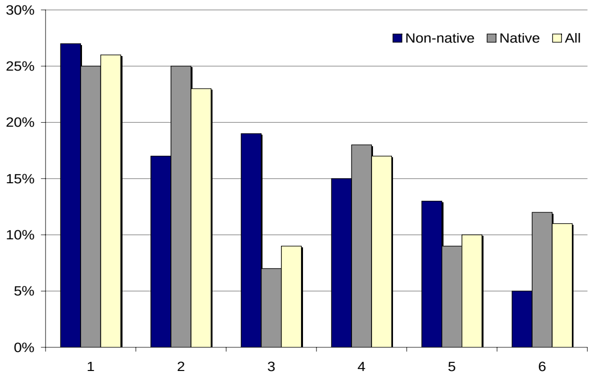

Figure 1: *Ranking for UK English speaker selection. Horizontal axis shows the speaker's identification number and the vertical axis indicates the relative preference according with mother tongue.*

An example of the variance among native and non-native speakers is presented in figure 1. In this case the depicted results are for the selection of a UK English speaker out of six candidates [9]. In most cases, the non-native listeners also preferred the candidate which received the highest rank by the native listeners. In some cases, the opinions between younger and older natives differed more than between natives and non-natives. A similar behavior was observed for the other languages.

In figure 2 we can observe how the listeners' gender can influence the voice judgment. Some of the candidates are equally preferred by both genders but others, however, are clearly discriminated by women. Nevertheless the preferred voices show a more balanced score for both genders.

Without forgetting the described issues the target voices were evaluated according with the following subjective parameters: pleasantness (PLS), intelligibility (INT), sensuality (FEM), emotiveness (EMO), character (CHR) and speaking rate (SPD). Three more questions were asked addressing the listeners' judgment on the suitability of those voices on typical TTS application, namely e-mail, news and instructions reading. A 5 points rating scale was used, which means that all voices were classified with marks from 1 (bad) to 5 (excellent) in every subjective attribute.

3.2. Objective Parameters

To objectively evaluate each of the recorded voices the following acoustic parameters were considered: F0 (mean, maximum,

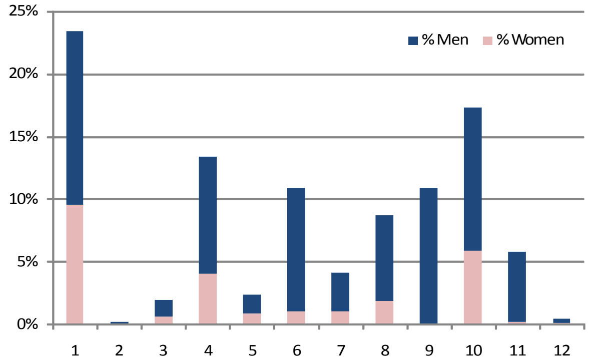

Figure 2: *Ranking for EP speaker selection. Horizontal axis shows the speaker's identification number and the vertical axis indicates the relative score according with listeners' gender. The shown were normalized to remove any bias resulting from the difference between the number of votes by gender.*

minimum, range and standard deviation), energy (mean and standard deviation), speaking rate (SPR in words per minute excluding pauses) and pausing rate (PAR) (rating between the duration of pauses and the total phonation duration without pauses). The features were extracted using Praat [11] and Mathworks Matlab.

Each parameter was independently correlated with the subjective evaluation results for finding acoustical clues of voice preference. The correlation values were calculated according to the equation:

$$Correl(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{(x-\bar{x})^2(y-\bar{y})^2}} \quad (1)$$

where X represents a set of x values and $\bar{x}$ their average, the same applies to Y. The output values are in the range -1 to 1, with 1 indicating a high linear relationship between the sets, -1 the inverse and 0 meaning that there is no linear dependence between sets.

4. Results and Discussion

The subjective evaluation results were used to build an ordered candidate ranking for each language but, alone, are useless for establishing direct comparisons of speakers or for any cross-language analysis. The same happens with the objective results. After gathering the results of subjective and objective assessments a joint analysis was performed.

The speaking rate analysis results are presented in figure 3. Again EP+BP and the other languages are presented separately. It can be observed that a high speaking rate is a desired characteristic for all the languages. The flat lines around 2.7/2.8 words per second seem to indicate that this an interesting value for this characteristic and that values higher than this can decrease the voice score. The multi-lingual analysis of this parameter can be misleading because same languages have much longer words (for example in German that has agglutinative processes in words composition) than others.

In figure 4 we show the mean fundamental frequency (F0) ordered by candidate ranking (1 is the best scored voice and 5 is the worst scored voice for this speaker sample). The results and presented separately for EP plus BP (darker line) and for all the other languages (lighter line). We can see that the analyzed f0 averages fit within a 20Hz range (around 26Hz on Mel perceptual scale) but yet presents a good diversity. The interesting observation is that, despite EP plus BP voices present lower F0 values and the other voices present higher F0 values, they all converge to a same common frequency band around 193Hz. This indicates not only a gold value for the fundamental frequency of a female voice but also the cross-lingual uniformity of this finding.

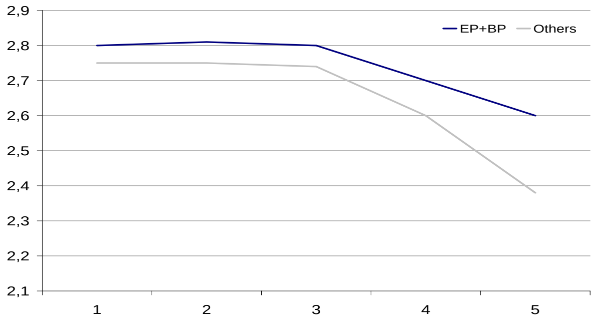

Figure 3: *Average relative scores per language according to fundamental frequency related parameters. Horizontal axis shows linear frequency (Hz).*

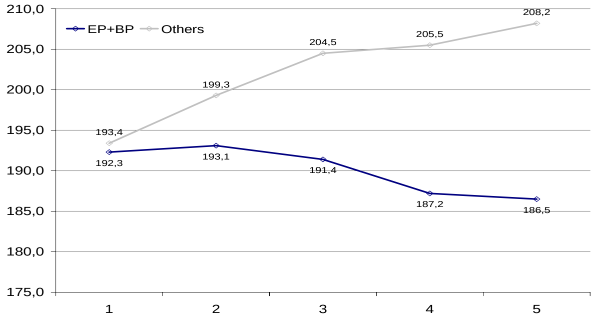

Figure 4: *F0 mean for the first five best scored voices. Horizontal axis shows the candidate ranking and vertical axis shows linear frequency (Hz).*

Still concerning the fundamental frequency, we can see in figure 5 a dot cloud on a fundamental frequency versus relative score plane. Each point represents the f0 for the best ranked voice for a given language and has an associated score. We can observe three clusters for minimum, average and maximum f0 values with increasing spatial variance. For the maximum f0 there is a trend for increased rankings on higher frequencies. This indicates that despite the low f0 preference it is also desirable to have a good vocal dynamic. The minimum f0 frequencies show a very small variance and the preferred values are close to the cluster f0 values. The average f0 cluster has a triangular shape with the best scores given to the lowest frequency values.

On another analysis we tried to understand what perceptual strategy is, mostly unconsciously, used by the listeners to evaluate subjective parameters. In table 1 we show, for a joint analysis of EP plus BP, the correlation values between the voices' scores for each subjective parameter and the related objective parameters. Fundamental frequency seems to be an important

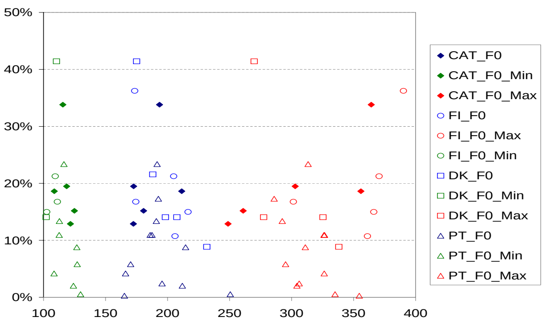

Figure 5: *Relative scores according to f0 values. Average, minimum and maximum values are presented for each language. Horizontal axis shows linear frequency (Hz) and vertical axis shows relative score.*

parameter on the evaluation of voice quality and it is also useful for judging character. We can also observe that both jitter and shimmer have negative correlation. This could be an expected observation but can point that high jitter specially reduces the perception of an emotional voice and that high shimmer contributes to decrease intelligibility. Harmonic to noise ratio is also inversely related with speaking rate. This may mean that when a high dynamic is imposed to the phonatory system the capacity to procedure harmonic sounds is reduced because tissues settling time has a longer relative duration for each sound. Pitch has the higher correlation values which emphasize its importance on the judgment of a voice.

Table 1: *Correlation between objective and subjective scores for EP plus BP.*

	PLS	INT	FEM	EMO	CHR	SPD
SPR	-0,13	-0,30	0,26	0,30	0,24	-0,01
Jitter	-0,62	-0,50	-0,76	-0,89	-0,79	-0,06
Shimmer	-0,58	-0,90	-0,37	-0,07	-0,30	-0.83
HNR	-0,77	-0,71	-0,70	-0,41	-0,65	-0,95
Pitch	0,97	0,50	0,29	0,78	0,84	0,57

An identical correlation table was produced for the remaining languages and the absolute differences with table 1 are presented in table 2. We can observe that all the values are below 0.20 and that 30% of the values are below or equal to 0.10. This may indicate that the obtained results for EP and BP voices and judgments are coherent with the values for other European languages. A statistical analysis using a z-test came to confirm that this is a valid assumption for a 90% confidence interval.

Table 2: *Absolute difference between two sets of correlation values for EP plus BP and the other analyzed languages.*

	PLS	INT	FEM	EMO	CHR	SPD
SPR	0,02	0,09	0,00	0,14	0,04	0,12
Jitter	0,14	0,04	0,13	0,20	0,19	0,11
Shimmer	0,11	0,16	0,06	0,03	0,16	0,17
HNR	0,11	0,00	0,15	0,20	0,12	0,14
Pitch	0,17	0,11	0,15	0,16	0,04	0,02

5. Conclusions

In this paper we described the construction of a multi-lingual multi-speaker voice database for voice quality analysis. The collected voices were evaluated by human listeners according with a set of subjective parameters that allowed creating a voice preference ranking. Additionally a set of objective parameters were extracted and correlated with each individual rank. This joint analysis leaded to several new interesting conclusions. Mainly we showed that the fundamental frequency values that gather more preferences are around 193Hz and that a speaking rate of 2.7/2.8 also brings additional votes. These results were also analyzed in two groups: one with the Portuguese language (European and Brazilian varieties) and another with a set of 7 European languages. We showed that the preference for an EP or BP female voice is perceptually identical to the preferences found on other European voices in distinct languages.

The correlation results between objective and subjective parameters are still preliminary but it was shown that there is a correlation between the voice quality ranking obtained by subjective listening tests and acoustic parameters. These parameters can therefore be used for an automatic preselection of promising speakers from a larger number of candidates. Further investigations will focus on the results obtained by different groups of listeners as young/old, native/non-native, and expert/non-expert regarding the speech processing technology. A more comprehensive analysis evolving all the languages is on going and will be published on future work.

6. References

[1] S. Pinho and P. Pontes, “Escala de avaliação perceptiva da fonte glótica: Rasat,” *Vox Brasilis*, vol. 8, no. 3, pp. 8–13, 2002.

[2] J. Kreiman, B. Gerratt, G. Kempster, and A. Erman, “Perceptual evaluation of voice quality. review, tutorial and a framework for future research,” *Journal of Speech and Hearing Research*, vol. 36, pp. 21–40, 1993.

[3] L. Eskenazi, D. G. Childers, and D. M. Hicks, “Acoustic correlates of vocal quality,” *Journal of Speech and Hearing Research*, vol. 33, pp. 298–306, 1990.

[4] J. Kreiman and B. R. Gerratt, “Sources of listener disagreement in voice quality assessment,” *Journal of Acoustical Society of America*, vol. 108, no. 4, pp. 1867–1876, 2000.

[5] S. Blaustein and B. Asher, “Reliability of perceptual voice assessment,” *Journal of Communications Disorders*, vol. 16, pp. 157–161, 1983.

[6] C. de Bruijn and S. Whiteside, “Effect of experience levels on voice quality ratings,” in *Proc. of Phonetics Teaching and Learning Conference*, London, August 2007.

[7] 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.

[8] D. Braga, L. Coelho, F. G. R. Junior, and M. S. Dias, “Subjective and objective evaluation of brazilian portuguese tts voice font quality,” in *Proc. of 14th International Workshop on Advances in Speech Technology*, Maribor, Slovenia, July 2007, pp. 306–311.

[9] O. Jokisch, G. Strecha, and H. Ding., “Multilingual speaker selection for creating a speech synthesis database,” in *Proc. of Workshop Advances in Speech Technology AST*, Maribor, Slovenia, 2004.

[10] J. Kreiman, B. Gerratt, and K. Precoda, “Listener experience and perception of voice quality,” *Journal of Speech and Hearing Research*, vol. 33, pp. 103–115, 1990.

[11] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (version 5.1.05),” *http://www.praat.org/*, 2009.