Subjective and Objective Assessment of TTS Voice Font Quality

Daniela Braga[1], Luís Coelho[2], Fernando Gil V. Resende Jr[3] and Miguel Sales Dias

[1] Microsoft Language Development Center, MLDC, Portugal

[2] Polytechnic Institute of Oporto, Portugal

[3] Federal University of Rio de Janeiro, Brazil

Abstract

In this paper, it is our aim to define a set of objective acoustic criteria, based on subjective listeners assessment of talent voices, that can help to rate the voice font quality, bearing in mind the development of Text-to-Speech systems (hereafter TTS). The voice talent selection process is seldom discussed in scientific papers, since academic community is obviously more concerned with synthetic voice evaluation than with modal voice talent assessment. However, the choice of a voice that will be used in TTS systems is definitely of extreme importance. In this work, we describe the selection process of the European Portuguese voice talent, based on socio and linguistic criteria and supported by subjective and objective tests and measurements of modal voice quality. Relationship between subjective and objective parameters is presented and discussed. A subjective test comparing natural and resynthesized voices was also conducted and its results are discussed.

1. Introduction

In the last few years, TTS technology has been widely improved in several aspects. The speech is more fluent and continuous, the intelligibility is higher, even emotions can be now synthesized. These developments lead to a real naturalness and created the new challenge of starting embedding TTS systems with working objects and tools, such as computers, cellular phones, Personal Digital Assistants (PDAs) and other devices, in order to create a voice help that is always present in daily human tasks. This permanent contact with a voice should be reliable and pleasant. But how good is a voice? Which acoustic and perceptive parameters make people prefer one voice to another? Is it possible to measure voice pleasantness? Questions like these together with the development of a new voice font for European Portuguese (EP) TTS systems are in the basis of the present study. In specialized literature, voice quality is often associated with voice impairments or disorders. There is an extensive bibliography on this subject, in which voice quality means normal speech characteristics that are somehow affected for pathological reasons, leading to hoarseness, roughness, raspiness, effort to talk, breathiness, vocal fry, uncomfortable or abnormal pitch, and other abnormal vocal symptoms [1], [2], [3], [4]. Still related with the clinical field and from a physiological point of view, voice quality is also used in phonetic and acoustic descriptions, such as the one carried out by Laver [5]. In this paper, the author describes the dynamics of six basic phonation types (*modal voice* – the so-called "normal voice"*, falsetto, whisper, creak, harshness and breathiness*) and possibilities of combination of each one's settings, generating compound phonation types (such as *whispery falsetto*, *whispery creak*, *harsh whispery voice* and so on). This author encompasses both supralaryngeal and laryngeal phonatory settings. This labelling proposal has been widely applied to speech analysis and synthesis. Another context in which voice quality is often used is the evaluation of telephone speech output [6] and synthesized speech. Many papers address several overall subjective tests, like the Mean Opinion Score (MOS) test [7], or intelligibility tests such as Semantically Unpredictable Sentences (SUS) test, the Diagnostic Rhyme Test (DRT), amongst many other tests (for a review on speech quality tests in TTS, see [8]). Voice quality is also associated with prosody generation and synthesis of emotions. Some authors explore the role of voice quality (meaning in this context fundamental frequency (hereafter F0), intensity, tempo aspects and spectral parameters) in emotion, mood and attitude expression [9], [10]. Campbell et al., within the

same framework, call it "the 4th prosodic dimension" [11].However, the assessment of modal voice quality (in Laver's expression, explained above [5]) is rarely discussed. An interesting study addressing this subject was conducted by Syrdal et al. [12] in order to check the suitability of a speaker's 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 explore the correlation between acoustic characteristics (RMS energy, breathiness, long-term spectra, f0, formants and bandwidths, speaking rate, concatenation and target costs) and the subjective attributes of synthetic speech quality (intelligibility, naturalness and pleasantness). Nevertheless, there is not much description on the speakers' selection process, which is said to have been made "empirically", although all the candidates (6 females and 9 males) were said to be professional speakers. The lack of studies in the area of subjective and objective criteria for voice talent selection opens the possibility for the research presented in this paper. This paper is structured as following: in section 2, the voice talent selection process that led to the identification of the TTS voice talent is described; two conducted subjective tests and corresponding results are presented and discussed in section 3; in section 4, objective measurements are shown; in section 5, tests and results comparing natural and resynthesized voice samples are discussed; in section 6, main conclusions are pointed out and future work is foreseen.

2. Voice talent selection

The voice talent selection process was developed in four stages. In the first stage, a national call for voice talents was launched. A few mandatory profile requirements were defined, such as: being female, having EP as mother language, having studied in Portugal up to the university level, speaking standard European Portuguese (although other dialectal varieties were expected to be considered) and having preferably some radio or theatre vocal experience. From a total of 485 presented female candidates, 74 were admitted to the second stage of evaluation. At this stage, candidates were invited to send us samples of their voices, with no restrictions of duration, with the maximum quality they could produce (acquired by their own means, either in professional recording studios, from their portfolio, or from their own PC). A subjective test was conceived with 13 questions, based in the MOS scale, and was conducted by 7 adult listeners (2 females and 5 males), familiarized with Speech Processing technology. The test was blind, that is to say each candidate and corresponding voice sample were given a number, in order to avoid any partial judgment and individually conducted. The listeners were asked to hear the voice samples (with headphones) and to rate them according to a set of previously defined subjective attributes (section 3). After this test, the 12 best scored candidates were invited to record a small text selected by us (of 219 words, phonetically and prosodically rich, with emotion expressions) in a professional recording studio (third stage of evaluation). Our goal with this casting was producing a final survey in which all the voices would be evaluated under the same conditions of sampling rate and text type. We produced a subjective test survey with 13 questions, assessing essentially the same attributes as in the previous subjective test, but which were differently rated. In other words, the first test had a 5 points rating scale, whereas the second test was an exclusive multiple choice questionnaire where only the best voice for each attribute could be selected. The survey was carried out by 74 listeners (19 females and 55 males) that were not familiarized with Speech Processing technology (third stage of evaluation). The voice ranking was obtained through the sum of votes each voice received along the survey. The In the fourth stage of evaluation, an objective analysis was carried out to confirm evidence provided by the subjective tests (section 4).

3. Subjective tests and results

In the 2nd stage of evaluation, Test 1 was conducted. 74 voice samples of up to 2 minutes recording time were evaluated according with the following subjective parameters: pleasantness (PLS), intelligibility (INT), articulation (ART), accent pronunciation (ACP), expressiveness (EXP), exceptionalness (EXC), sensuality (SNS), attitude (ATT). Three more questions were asked addressing the listeners' judgment on the suitability of those voices for the following applications: e-mail, news or instructions reading. This was a 5 points rating scale, which means that all voices were classified with marks from 1 (bad) to 5 (excellent) in every subjective attribute. In Test 1 results show that: the most pleasant voices were 7 (rated 4.7), 11 and 12 (both rated 4.6); the sexiest voices were 7 and 11 (rated 4.6); voice 7 showed more attitude and determination (rated 4.6); voice 11 seemed to be more special and hard to forget (rated 4.4); concerning articulation, all voices were well classified, but the best were 3, 4, 5, 6, 7 and 12, (rated 4.6); regarding accent pronunciation and intelligibility, all the 12 best scored voices showed no dialectal mark, being all classified over 4.4; voice 10 was judged as the most expressive one (rated 4.7). Concerning voice applications, voice 11 received the highest preference for e-mail (rated 4.7) and news reading (rated 4.3); although for instructions voice 7 was preferred.

For age assessment, listeners were asked to guess voices' age in the following intervals: 1 – under 25 years old), 2 – between 25 and 30, 3 – between 30 and 35, 4 between 35 and 40, 5 – over 40. All 12 finalist voices were perceived to be between 2.4 and 3.3 points, which means that the preferred perceptual voice age rounded 27 and 31.5 years old. This evidence is interesting when analyzing age perception of the less scored candidates, whose age was perceived as being over 35 years old. Concerning speaking rate, all 12 best scored voices were considered normal, neither too fast nor too slow. Another interesting issue is that perceived age does not correspond to real age, since most of the 12 final speakers are over 31. Although we have used most of the MOS scale parameters, we have excluded naturalness, comprehension and listening effort. In fact, these parameters are crucial in synthetic voice quality, but are redundant when assessing normal and professional modal voices that are recorded in good quality conditions.

In the 3rd evaluation stage, Test 2 was carried out with the purpose of deciding the best Portuguese voice that will be integrated in our EP TTS system. Therefore, the test could not be too much time consuming, in order to be able to be carried out by as many listeners as possible. That is why we excluded some subjective parameters that were assessed in Test 1, such as EXC and ACP. INT and ART were assessed in the same question. The test was designed as a survey. First questions were on listener's gender and mother language. Middle questions were on the PLS, SNS, ATT, INT, SPR and voice applications (e-mail, news and instructions reading). Last question was on age perception of the preferred voice. Test 2 was rated according with the listeners' best choice for each question.

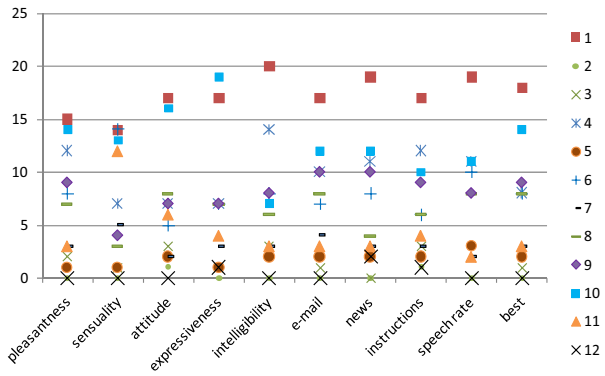

Fig.1. Test 2 results for the 12 best scored voices.

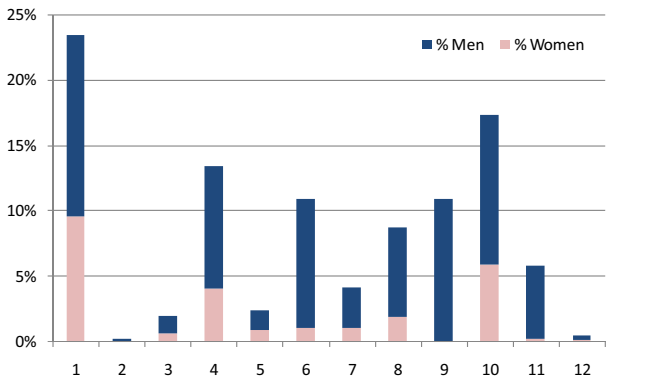

Fig. 3. Test 2 overall results for the 12 best scored voices according with listeners' gender.

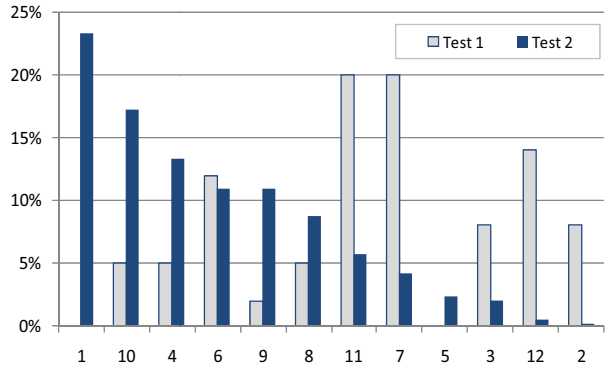

Fig. 2. Test 1 and 2 overall results for the 12 best scored voices

	PLS	SNS	ATT	EXP	INT	EMA	NEW	INS	SPR	BES
PLS	1,00	0,72	0,91	0,92	0,89	0,97	0,95	0,95	0,95	0,96
SNS		1,00	0,73	0,76	0,63	0,68	0,72	0,66	0,71	0,73
ATT			1,00	0,97	0,77	0,91	0,87	0,88	0,87	0,95
EXP				1,00	0,74	0,91	0,89	0,85	0,87	0,96
INT					1,00	0,91	0,93	0,97	0,94	0,87
EMA						1,00	0,96	0,97	0,97	0,98
NEW							1,00	0,97	0,96	0,95
INS								1,00	0,95	0,93
SPR									1,00	0,96
BES										1,00
TOT										

Table 1. Subjective parameters correlation.

In Figure 1, the assessed subjective parameters are displayed. Voice 1 occupies the first place in most of the subjective parameters, except in EXP, in which Voice 10 is best scored. The "best" parameter in Figure 1 corresponds to the question: "In your opinion, what is the best voice in general?". In Figure 2, the Test 1 and 2 overall ranking of the 12 best scored voices is displayed. This chart shows very surprising results, since the voice ranking has dramatically changed from Test 1 to Test 2. In Test 1, the best scored voices were voices 7 and 11, but in Test 2, the best scored voices were 1 (23%), followed by voices 10 (17%), 4 (13%) and 6 (11%) some of the worst scored voices in Test 1. A possible explanation for these results is the content of voice samples used in Test 1. Sample for Voice 11, for instance, was a popular TV spot advertising a prestigious car brand with a sophisticated music behind. Many samples were commercial spots or parts of documentaries. Another fact that may have influenced the first group of listeners was the presence of image data, as some voice samples were video clips. To prevent these variation factors in Test 2, the voice content was a previously provided text to all candidates and only voice was recorded, under the same studio conditions.

The results were totally opposite as firstly expected. Voice 1 is positioned in the best place in most of the considered subjective attributes, immediately followed by Voice 10, also better positioned than in Test 1. These results seem to confirm the importance of this methodology in voice talent selection process. Despite the doubtless victory of Voice 1, Figure 3 shows different voice quality preferences between men and women. Male listeners' preference by Voice 1 (14%) is not so far from Voices 9 and 10 (11%) or even Voice 6 (10%) and 4 (9%), which means that men's preferences are more varied. But when looking at women's preferences, we can see that voices that both consider being more sensual (Voices 6, 9, 11) are not much in women's preferences. Furthermore, women seem to prefer dynamic and professional voices, rather than sexy voices. In Table 1, a subjective parameters correlation matrix is drawn, using the following equation:

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

Some interesting conclusions can be obtained: a pleasant voice (PLS) is strongly correlated with an assertive (ATT) and expressive (EXP) voice. The best voice is mainly pleasant (PLS), expressive (EXP), fast (SPR) and excellent for e-mail reading. Sensuality (SNS) and intelligibility (INT) are less important attributes in a good voice quality. Sensuality is less important in voice applications than pleasantness.

Concerning age assessment, listeners seem to prefer voices sounding between 26 and 30 years old. 60.8% of the listeners classified their preferred voice in that age range, followed by 27% that selected a voice in the 31-35 age range. From the total number of people voting in Voice 1, 83.3% classified it in the 26-30 range, which shows a much defined age perception for this voice. On the contrary, Voice 10, secondly placed, showed dispersion in age perception: 46.1% of listeners rated it in the 26-31 range, 30.7% in the 31-35 range and 23.0% in the 20-25 range.

Table 2. F0 analysis

Voice	% Class.	Mean	Min	Max	Range	St. Dev.
1	23	206,6	116,4	345,8	229,4	58,0
2	0	165,2	83,3	354,5	271,2	59,0
3	2	211,9	124,1	304,3	180,2	57,1
4	13	198,4	78,9	288,8	209,8	38,1
5	2	195,6	96,2	306,1	209,9	57,6
6	11	186,0	112,6	326,6	214,0	61,2
7	4	166,1	108,5	326,3	217,8	45,8
8	9	214,5	126,8	311,0	184,2	54,7
9	11	187,4	76,4	326,1	249,7	59,0
10	17	190,5	76,0	285,9	210,0	49,6
11	6	170,2	127,1	295,1	168,0	45,5
12	1	250,5	129,9	335,0	205,2	55,4

Table 3. Energy, SPR and PAR analysis

		Normalized Energy		Speaking Rate	Pause Rate
Voice	%Class.	Mean	St. Dev.	(words/s)	(pause/tot)
1	23	0,081	0,145	2,80	37,25
2	0	0,112	0,201	2,45	36,78
3	2	0,104	0,163	2,64	30,13
4	13	0,063	0,139	2,43	45,16
5	2	0,105	0,197	2,52	37,34
6	11	0,158	0,242	2,82	42,74
7	4	0,099	0,167	2,22	42,49
8	9	0,050	0,100	2,64	44,67
9	11	0,100	0,184	2,88	34,62
10	17	0,123	0,189	2,81	38,37
11	6	0,059	0,138	2,50	45,41
12	1	0,078	0,134	2,56	28,88

4. Objective tests and results

The following acoustic parameters were assessed in fourth stage of evaluation: F0 (mean, maximum, minimum, range and standard deviation), energy (mean and standard deviation), speaking rate (SPR in words per minute excluding pauses) and pausing rate (PAR) (total duration of voice sample without pauses). In Tables 2 and 3, F0 analysis, energy, speaking rate and pausing are displayed for each voice. Tables' inspection show that listeners prefer: 1) female voices with mean F0 ranging between 186 and 206 Hz and seem to dislike mean F0 ranges under and over those values (e.g. Voices 2 and 12 with very low and very high mean F0s, respectively); 2) voices with a high speaking rate combined with long utterance breaks. Energy seems to be a parameter with little influence in voice quality judgment. Data displayed in Table 4 allows us to confirm the importance of a voice with a high speaking rate combined with well-defined pausing and shows that listeners seem to prefer voices with low minimum F0s. The correlation equation (1) was also used.

5. Resynthesis tests

Based on some authors who state that the perceived quality of a natural voice may not predict its synthetic quality [12], a different subjective test with the resynthesized voices was also conducted. Our goal with this test was assessing the 12 natural voices used in Test 2 after resynthesis. The chosen manipulated parameters in resynthesis were F0 and durations. F0 was set constant to the initial mean value. Energy was normalized. Durations were changed with PSOLA algorithm. Two sets of voice samples were presented to 8 different listeners who were not familiarized with TTS technology. The most interesting conclusion drawn from this test is the consensual opinion of Voice 10's good quality, which was secondly scored in natural and resynthesized voice quality rankings.

Table 4. Acoustic features correlation with voices score.

Feature	Correlation
f0_mean	0,014
f0_min	-0,209
f0_max	-0,115
f0_range	0,067
f0_stdev	-0,099
Int_mean	-0,014
Int_stdev	-0,061
Speech Rate	0,538
Pause Rate	0,269

6. Conclusions

In this paper, a methodology in four stages to assess TTS voice font quality is proposed. Subjective tests and objective analysis and correlations were described and results were discussed. The following main conclusions can be drawn from this work: 1) a good voice is subjectively pleasant, assertive, expressive and preferably between 26 and 30 years old; 2) the corresponding objective characteristics are F0 ranging between 186 and 206 Hz, low minimum F0s and high speaking rates combined with well-defined pausing. These results can be considered in synthetic speech generation. As future work, we intend to extract the spectral characteristics of first and second formants and first and second formant bandwidths and analyze their correlation with the subjective parameters. As an extension of this research, a similar study for Brazilian Portuguese voice quality assessment is ongoing. Based on our present experience, we propose the following changes in order to improve our methodology: 1) not allow video clip files as voice samples; 2) randomly change the order of voice samples every time a listener goes through the test or survey; 3) final selected voices should be no more than 8, in order to prevent judgment dispersion.

References

1. *R. J. Bake*, Clinical measurement of speech and voice. College Hill, Boston, 1987, pp. 197-240.

2. *J. Kreiman, B.R. Gerratt, G.B. Kempster, A. Erman*, "Perceptual Evaluation of Voice Quality. Review, tutorial and a framework for future research", J. Speech and Hearing Research, vol. 36, 1993, pp. 21-40.

3. *L. Eskenazi, D. G. Childers, D. M. Hicks*, "Acoustic Correlates of Vocal Quality", J. of Speech and Hearing Research, Vol.33, 1990, pp. 298-306.

4. *J. Kreiman, B. R. Gerratt*, "Sources of listener disagreement in voice quality assessment", J. Acoust. Soc. Amer., 108 (4), 2000, pp. 1867-1876.

5. *J. Laver*, The Phonetic Description of Voice Quality, Cambridge University Press, Cambridge, 1980.

6. ITU-T Recommendation P.85, Telephone transmission quality subjective opinion tests. A method for subjective performance assessment of the quality of speech voice output devices.

7. *M. Viswanathan, M. Viswanathan*, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale", Computer Speech and Language, vol. 19, January 2005, pp. 55-83.

8. *S. Lemmety*, Review on Speech Synthesis Technology, Master's Thesis, Helsinki University of Technology, 1999.

9. *C. Gobl, A. Chasaide*, "Testing affective correlates of voice quality through analysis and resynthesis", Proc. of the ISCA Workshop on Speech and Emotion, 2000, pp. 178-183.

10. *O. Turk, M. Schröder, B. Bozkurt, L. M. Arslan*, "Voice Quality Interpolation for Emotional Text-to-Speech Synthesis", Proc. of Interspeech 2005, Lisbon, Portugal.

11. *N. Campbell, P. Mokhtari*, "Voice Quality: the 4th Prosodic Dimension", 15th Intern. Congress of Phonetic Sciences, 2003, pp. 2417-2420.

12. *A. Syrdal, A. Conkie, Y. Stylianou*, "Exploration of Acoustic Correlates in Speaker Selection for Concatenative Synthesis", Proc. of ICSLP 98, 1998.