VOICE PLEASANTNESS: ON THE IMPROVEMENT OF TTS VOICE QUALITY

Luis Coelho[1], *Daniela Braga*[2], *Carmen Garcia-Mateo*[3]

[1] Instituto Politécnico do Porto, ESEIG, Porto, Portugal
[2] MLDC - Microsoft Language Development Center, Lisbon, Portugal
[3] Universidad de Vigo, Dpto. Teoria de la Señal e Telecomunicaciones, Vigo, Spain

RESUMEN

The aim of this paper is to validate the objective description of the voice pleasantness concept. This concept has been objectively defined on an exhaustive study based on subjective voice analysis, comparison and scoring[1, 2]. For increasing text-to-speech (TTS) voice pleasantness while maintaining speaker's identity a set of procedures, that operate on time and frequency domains and encompass phonetic and prosodic levels, are proposed. These enhancement procedures have been integrated in widely know speech engines and the obtained results showed an effective gain on the evaluated parameters providing support to both methodology and objectives. The Mean Opinion Score (MOS) performance evaluations indicated pleasantness improvements of 7% for female voices and 3% for male voices when compared with systems that do not consider these quality aspects.

1. INTRODUCTION

The main goal on speech synthesis systems is the production of high quality speech with the maximum naturalness. The latest developments in acoustic engines are slowly fulfilling these requirements and TTS technology is increasingly being included in our daily lives. Front-end design has also improved and can now deliver rich prosodic information to the acoustic engine which efficiently produces human like speech. Our concerns regarding voice enhancement and quality assessment are often associated, in specialized literature, with voice impairments or disorders [3, 4]. However for a continuous use of TTS systems on a daily basis additional quality requirements arise. On another perspective, considering that no voice pathologies are present, the interaction with a machine's voice must be everything but boring and if possible it should be pleasant. It is our belief that one of the next frontiers on TTS technology resides on improving voice quality in order to help to increase interaction pleasantness.

The rest of the paper is organised as follows: In section 2 we describe the used methodology for processing the speech signal. In section 3, a quick presentation is made on what are parameters can be used to define pleasantness and how the quality goal is obtained. In section 4 we show several performance evaluation tests and an extensive discussion of the results. Finally, in section 5, main conclusions are pointed out and future work is foreseen.

2. METHODOLOGY

To validate the objective pleasantness definition we propose a set of adaptive phone level signal operations for voice enhancement according to the pleasantness parameters. These procedures can be easily integrated with an existent system as an extension, independently of the technology, or can be directly included as a part of the processing operations. Traditional voice conversion techniques [5, 6] involve intense signal manipulation which often lead to quality degradation. In our case quality is a main concern and speaker's identity must be preserved so we decided to support our signal operations with mature proven models.

2.1. Integration with TTS as an output module

Signal processing operations will be performed on the TTS' output speech. The required information consists of phoneme sequence, timing, voiced/unvoiced and prosodic tags with pitch, all given by the front-end, and additionally an enhancement goal in percentage, for each parameter.

The time domain signal decomposition/composition is based on the Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) algorithm [7] since it introduces minimal distortions and is a very effective way to perform small pitch changes. Segment durations are changed accordind to a context and phoneme dependent ratio:

$$d_{new}(p) = \frac{1}{2C+1} \sum_{i=p-C}^{p+C} \frac{d_{target}(i)}{d_{cur}(i)} \qquad (1)$$

where d_{new} is the new duration for segment p, d_{ref} and d_{cur} are the target and current durations respectively and C is a neighbourhod limiter for analysis. In our case most considerer segments are phonemes and a neighborhood of 2 segments, before and after, was used ($C = 2$).

The frequency domain operations are performed by a filter that is calculated for each window considering the current phoneme. The filter is defined according to the following procedure:

$$s(k) = \sum_i a_i s(k-i) + G.u(k) \tag{2}$$

is adjusted to a source-filter model based on an all-pole filter

$$H(z) = \frac{G}{1 - \sum_{i=1}^{N} a_i z^{-i}} = \frac{G'}{\sum_j (z^{-1} - p_j)} \tag{3}$$

with $U(z)$ as the excitation, G as the model gain and p_j as filter poles. For a sufficient model order N the signal can be successfully reconstructed since both filter and excitation information are known. In our case the modified covariance method [8] was used for Auto-Regressive (AR) estimation, because it leads to stable models, it can efficiently deal with wrong model orders and because it introduces reduced spectral line splitting effects. 2. Extract formant frequencies and bandwidths. For each filter pole p_j with magnitude A_j and frequency w_j, the formant frequency F_j, for a sampling frequency f_s comes as:

$$F_j = f_s \frac{w_j}{2\pi} \tag{4}$$

The bandwidth B_j are approximated by:

$$B_j = f_s \frac{A_j}{\pi} \tag{5}$$

3. Adjust formant frequencies by relocating filter poles and perform formant sharpening. Each (F_j, B_j) pair is then manipulated to fit the given context dependent criteria.

With all this it is possible to build the correction filter that will be applied to each specific window. A discrete Itakura-Saito based metric is used to evaluate the spectral distance between original and enhanced signal:

$$d_{IS} = \frac{1}{N} \sum_{m=1}^{N} \left(\frac{P(\omega_m)}{\hat{P}(\omega_m)} - \log \frac{P(\omega_m)}{\hat{P}(\omega_m)} \right) \tag{6}$$

A threshold limits the extent of spectral changes in the signal.

4. Spectral weighting. Finally, for reducing the effects of noise amplification introduced by the correction filter, an extra spectral weighting filter is used. The filter is based on Linear Prediction (LP) coefficients using a zero/pole transfer function with coefficient weighting:

$$W(z) = \frac{A(z/\alpha)}{A(z/\beta)} = \frac{1 + \sum_k a^k z^{-k}}{1 + \sum_l a^l z^{-l}} \tag{7}$$

The values for α and β are computed for each window according to the previously extracted spectral envelope and, to keep a low distortion between sudden changes on the parameters, an interpolation is performed considering the previous values. Initial values are found by first adjusting α and then β, in successive iteration, with spectral distance restrictions. Tipical values are $\alpha \in [0.85, 1.00]$ and $\beta \in [0.82, 0.95]$.

2.2. Integration with HMM based TTS

We also applied the voice enhancement criteria to an HMM based synthesis system [9]. Since these systems are centered on time-frequency statistical sound models, the voice changes can be performed by small arrangements on the model's statistical parameters. In our case we trained a voice font with what we considered to be an optimal voice (our optimal voice resulted from a set of several subjective tests from which the enhancement parameters were derived) and using a weighted re-estimation process, based on the Baum-Welch algorithm, we adapted the desired voice font. The binary trees that help to determine the synthesis parameters were also rebuilt. All the process was conducted during the training and no changes were made to the speech production engine.

3. VOICE QUALITY

The studies performed on [1] and [2] correlate objective parameters such as formant frequencies, fundamental frequency, duration, etc. with with subjective scores. In an unpublished work of the authors the study is extended to European Spanish, French and English with some preliminary results presented here. The subjective classification was performed on a five points scale (1 for the worst opinion and 5 for the best opinion) and the correlation values were obtained using the standard correlation equation:

$$correl(X,Y) = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^N (x_i - \bar{x})^2 \sum_i^N (y_i - \bar{y})^2}} \tag{8}$$

where $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_N\}$ are the set of objective results and the subjective scores respectively.

The objective evaluation for each voice and for each group show different results but share some common indicators. The results for the most representative parameters, in all the five languages, are presented averaged in table 1. The fundamental frequency is analysed considering the average, maximum and minimum, standard deviation and the difference between maximum and minimum values. Speaking rate (SPR) is also considered as the number of words per time unit (in this case words per second) and in percentage as the relation between the speaking time and non-speaking/pause time (a pause was defined as any non-speaking segment with duration greater than 20ms). The energy and the intensity were normalized before comparison.

The obtained values can be biased if a group of speakers share some similar voice characteristic. To avoid this

	Score Correlation	
Objective Feature	Average	St. Dev.
F0 Average (Hz)	- 0.56	0.052
F0 Min (Hz)	- 0.45	0.042
F0 Max (Hz)	- 0.67	0.023
F0 Std. Dev (Hz)	0.29	0.015
F0 range (Hz)	0.10	0.012
SPR (words/sec)	0.17	0.011
SPR (%)	- 0.28	0.018
Pause Rate (%)	0.30	0.011
Total Energy (dB)	- 0.54	0.034
Energy St. Dev.	- 0.52	0.030
Avg. Intensity	- 0.59	0.038

Table 1. Correlation results between objective features and voice classification scores

		Frequency		Bandwidth	
Phon.	Formant	EP	BP	EP	BP
A	1	- 0.33	- 0.36	- 0.20	- 0.18
	2	- 0.31	- 0.30	- 0.18	- 0.19
6	1	- 0.41	- 0.57	0.30	0.36
	2	- 0.32	- 0.52	- 0.21	- 0.27
i	1	- 0.51	- 0.32	- 0.19	- 0.19
	2	- 0.27	- 0.28	- 0.12	- 0.11
o	1	- 0.01	0.04	0.27	0.29
	2	- 0.05	0.02	0.17	0.17
E	1	0.15	- 0.43	- 0.18	- 0.14
	2	- 0.09	- 0.11	- 0.21	- 0.19
U	1	- 0.47	- 0.45	- 0.26	- 0.26
	2	- 0.14	0.21	- 0.22	- 0.21

Table 2. Correlation results between formant frequencies and formant bandwidth in some vowels and voice classification scores for European Portuguese (EP) and Brazilian Portuguese (BP)

the selected speakers in our case formed a very heterogeneous group with very distinct voices and speaking styles. From the results the following pleasantness indicators can be pointed: *Low fundamental frequency*, we can see that F0 average showed a high inverse correlation with the obtained voice score which indicates a preference for voices with a lower pitch; *Dynamic prosody*, the voices with greater variation on the prosodic parameters (not all are shown) also seemed to be preferred. Variations on pitch, durations and intensity are valued characteristics from the pleasantness point of view; *High speaking rate*, the fast speakers showed to be prefered nevertheless it is desirable to have well defined pauses between utterances. We suppose that this shows assertiveness and transmits confidence to the human listener.

The pleasantness reference is found by following the methodology described in [1]. It requires a group of candidate speakers for voice-font recording, a set of listeners for evaluations and an application context. The initial set of voices should be heterogeneous but compatible with generic voice pleasantness values (this can vary according to culture, language, etc.). For EP and BP female speakers the objective pleasantness reference is defined as $f_0 = 208 \pm 9Hz$ (95% confidence interval), normalized energy $18.8dB/sec$, 3.8 words/sec for the speaking rate and 15.3% of the time should be used for pauses. The values were obtained by maximizing a quadratic polynomial used to model the relation between objective parameter and score. The model is obtained by minimizing the error for a quadratic curve fitting using each of the analysed parameters:

$$k = \sum_{i=1}^{N} \left[y_i - (b_0 + b_1 x_i + b_2 x_i^2) \right]^2 \qquad (9)$$

where y_i represents the obtained score and x_i represent the parameters in evaluation. The maximum of the obtained curve will point the best values for the parameter.

Signal manipulation is performed on time and spectral domains. In the time domain phoneme level durations are changed by a segment dependent fixed ratio which is obtained from the average segment durations on original and reference voice. If the original speaker has a natural speaking style the introduced changes on duration will not introduce ambiguity on speaker identification. In spectral domain similar ratios are previously calculated for formants and related bandwidths in voiced sounds. These ratios have been limited to $[0.9, 1.1]$ to limit spectral changes. In this domain no changes are made to unvoiced sounds.

4. RESULTS AND DISCUSSION

We have used two EP voice databases, one with a female voice (87 min.) and another with a male voice (100 min.). In spite of all the careful taken with recording scripts and recordings no special attention has been given to speaker selection (there was only one person only good articulation and good dynamics at a subjective level were considered). The voices were objectively analysed and a pleasant voice model was built. The several considered parameters were then used to define a voice goal as described. A set of 10 long sentences (expected to give at least 10 seconds of speech) were chosen and the related speech was produced by a concatenative system based on Festival (CF) and by HTS (HTS) basic configuration, the same sentences were the produced using our proposed voice enhancement procedure (CFVE and HTSVE). A group of 11 listeners with no special speech technology knowledge evaluated the sentences and voted for voice pleasantness according to their preference. In figure 1 we show the results obtained in our subjective evaluation tests. It can be observed that the voice enhancement procedures introduce some improvements on the listener's opinion which means that the signal manipulation doesn't introduce any adverse effects. However the changes are not highly sig-

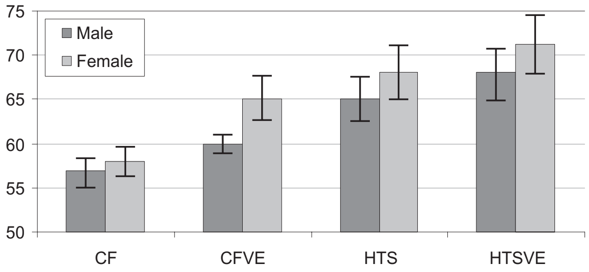

Figure 1. *Results of the subjective tests according to voice pleasantness (90% confidence interval is show in the top of the bars).*

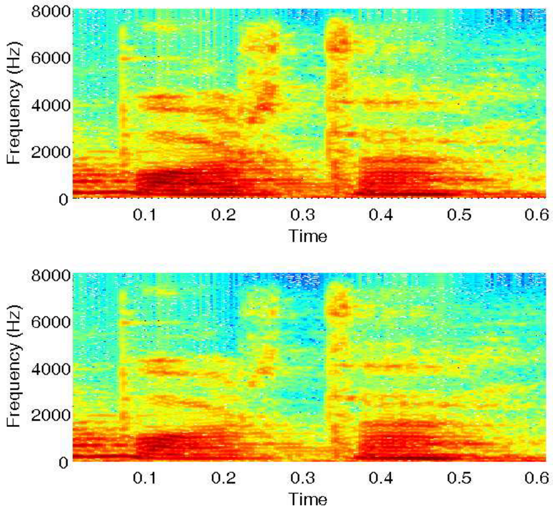

Figure 2. *Spectrum for the word "porta"(door) after (top) and before (bottom) voice enhancement for voice pleasantness.*

nificant because the speaker's identity was preserved. We expect that the results could show greater improvements if the original voice was a randomly selected voice.

For male voices only an average of 3% of improvements could be measured for both systems. For female speakers there is an increment of around 7% on voice pleasantness preference. The obtained results for HTSVE are not so expressive due to an insufficient amount of trainning material and due to the experienced issues on phoneme adaptation without losing speaker's identity. Nevertheless this suggests that the identified voice quality parameters are important for voice pleasantness and that it is possible to perform voice enhancement by the use of signal manipulation operations.

In figure 2, we can observe an example of the enhancement for a word pronounced by a female speaker. Looking at the spectrum represented above it is possible to see that the formants are better defined while the energy in the spectrum valleys remmains at the same levels, spectral peaks are also sharper. It is possible to see small changes in the durations of the vowels (largest segments).

5. CONCLUSIONS

In this paper, we presented a set of objective parameters that play an important role on the evaluation of voice quality, particularly concerning pleasantness. It is our belief that these parameters will be of major importance on voice selection and TTS systems' performance evaluations. After defining how to increase the voice quality we have proposed a set of procedures, considering time and frequency domain parameters, which can help to follow this path. We have performed some evaluation experiments and the obtained results show that our proposal can lead to an effective voice enhancement.

6. BIBLIOGRAPHY

[1] Daniela Braga, Luis Coelho, Fernando Gil Resende, y Miguel Dias, "Subjective and objective evaluation of brazilian portuguese tts voice font quality," in *Proc. of Advances in Speec Technology*, 2007.

[2] Daniela Braga, Luis Coelho, y Fernando Gil Resende, "Subjective and objective assessment of tts voice font quality," in *Proc. of SPECOM*, 2007.

[3] J. Kreiman, B.R. Gerratt, G.B. Kempster, y A. Erman, "Perceptual evaluation of voice quality. review, tutorial and a framework for future research," *Journal of Speech and Hearing Research*, pp. 21–40, 1993.

[4] J. Kreiman y 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] A. Kain y M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. of ICASSP*, 1998.

[6] Y. Stylianou, O. Capp'e, y E. Moulines, "Statistical methods for voice quality transformation," in *Proc. of Eurospeech*, 1995.

[7] E. Moulines y F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[8] S. M. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice-Hall, Englewood Cliffs, 1988.

[9] K. Tokuda, T. Yoshimura, T. Kobayashi, y T. Kitamura, "Speech parameter generation algorithms for hmm speech synthesis," in *Proc. of ICASSP*, 2000.