Automatic Evaluation of Synthetic Speech Quality by a System Based on Statistical Analysis

Jiří Přibil[1,2(✉)], Anna Přibilová[3], and Jindřich Matoušek[2]

[1] Institute of Measurement Science, SAS, Bratislava, Slovakia
Jiri.Pribil@savba.sk
[2] Faculty of Applied Sciences, Department of Cybernetics, UWB, Pilsen, Czech Republic
jmatouse@kky.zcu.cz
[3] FEE & IT, Institute of Electronics and Photonics, SUT in Bratislava, Bratislava, Slovakia
Anna.Pribilova@stuba.sk

Abstract. The paper describes a system for automatic evaluation of speech quality based on statistical analysis of differences in spectral properties, prosodic parameters, and time structuring within the speech signal. The proposed system was successfully tested in evaluation of sentences originating from male and female voices and produced by a speech synthesizer using the unit selection method with two different approaches to prosody manipulation. The experiments show necessity of all three types of speech features for obtaining correct, sharp, and stable results. A detailed analysis shows great influence of the number of statistical parameters on correctness and precision of the evaluated results. Larger size of the processed speech material has a positive impact on stability of the evaluation process. Final comparison documents basic correlation with the results obtained by the standard listening test.

Keywords: Listening test · Objective and subjective evaluation
Quality of synthetic speech · Statistical analysis

1 Introduction

At present, many objective and subjective criteria are used to evaluate quality of synthetic speech that can be produced by different synthesis methods implemented mainly in text-to-speech (TTS) systems. Practical representation of a subjective evaluation consists of a listener's choice from several alternatives (e.g.

The work was supported by the Czech Science Foundation GA16-04420S (J. Matoušek, J. Přibil), by the Grant Agency of the Slovak Academy of Sciences 2/0001/17 (J. Přibil), and by the Ministry of Education, Science, Research, and Sports of the Slovak Republic VEGA 1/0905/17 (A. Přibilová).

P. Sojka et al. (Eds.): TSD 2018, LNAI 11107, pp. 315–323, 2018.
https://doi.org/10.1007/978-3-030-00794-2_34

mean opinion score, recognition of emotion in speech, or age and gender recognition) or from two alternatives, speech corpus annotation, etc. [1]. Spectral as well as segmental features are mostly used in objective methods for evaluation of speech quality. Standard features for speaker identification or verification, as well as speaker age estimation, are mel frequency cepstral coefficients [2]. These segmental features usually form vectors fed to Gaussian mixture models [3,4] or support vector machines [5] or they can be evaluated by other statistical methods, e.g. analysis of variance (ANOVA) or hypothesis tests, etc. [6,7]. Deep neural networks can also be used for speech feature learning and classification [8]. However, they are not sufficient to render the way of phrase creation, prosody production by time-domain changes, speed of the utterance, etc. Consequently, supra-segmental features derived from time durations of voiced and unvoiced parts [9] must be included in the complex automatic system for evaluation of synthetic speech quality by comparison of two or more utterances synthesized by different TTS systems. Another application may be evaluation of degree of resemblance between the synthetic speech and the speech material of the corresponding original speaker whose voice the synthesis is based on.

The motivation of this work was to design, realize, and test the designed system for automatic evaluation of speech quality which could become a fully-fledged alternative to the standard subjective listening test. The function of the proposed system for automatic judgement of the synthetic speech signal quality in terms of its similarity with the original is described together with the experiments verifying its functionality and stability of the results. Finally, these results are compared with those of the listening tests performed in parallel.

2 Description of Proposed Automatic Evaluation System

The whole automatic evaluation process consists of two phases: at first, databases of spectral properties, prosodic parameters, and time duration relations (speech features – SPF) are built from the analysed male and female natural utterances and the synthetic ones generated by different methods of TTS synthesis, different synthesis parameters, etc. Then, separate calculations of the statistical parameters (STP) are made for each of the speakers and each of the types of speech features. The determined statistical parameters together with the SPF values are stored for next use in different databases depending on the used input signal ($DB_{ORIG}, DB_{SYNT1}, DB_{SYNT2}$) and the speaker (male/female). The second phase is represented by practical evaluation of the processed data: at first, the SPF values are analysed by the ANOVA statistics and the hypothesis probability assessment resulting from the Ansari-Bradley test (ASB) or the Wilcoxon test [10,11], and for each of their STPs the histogram of value occurrence is calculated. Subsequently, the root-mean-square (RMS) distances (D_{RMS}) between the histograms stemming from the natural speech signals and the synthesized ones are determined and used for further comparison by numerical matching. Applying the majority function on the partial results for each of SPF types and STP values, the final decision is got as shown in the block diagram in Fig. 1. It

is given by the proximity of the tested synthetic speech produced by the TTS system to the sentence uttered by the original speaker (values "1" or "2" for two evaluated types of the speech synthesis). If differences between majority percentage results derived from the STPs are not statistically significant for any type of the tested synthesis, the final decision is set to a value of "0". This objective evaluation result corresponds to the subjective listening test choice "*A sounds similar to B*" [1] with small or indiscernible differences.

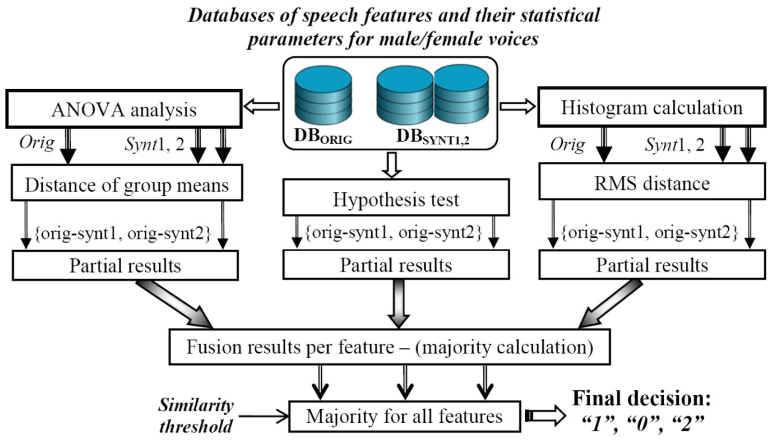

Fig. 1. Block diagram of the automatic evaluation system of the synthetic speech.

For building of SPF and STP databases, the speech signal is processed in weighted frames with the duration related to the speaker's mean fundamental frequency F0. Apart from the supra-segmental F0 and signal energy contours, the segmental parameters are determined in each frame of the input sentence. The smoothed spectral envelope and the power spectral density are computed for determination of the spectral features. The signal energy is calculated from the first cepstral coefficient c_0 (En_{c0}). Further, only voiced or unvoiced frames with the energy higher than the threshold En_{MIN} are processed to eliminate speech pauses in the starting and ending parts. It is very important for determination of the time duration features (TDUR). In general, three types of speech features are determined:

1. time durations of voiced/unvoiced parts in samples Lv, Lu for a speech signal with non-zero F0 and $En_{c0} \geq En_{MIN}$, their ratios Lv/u_L, Lv/u_R, Lv/u_{LR} calculated in the left context, right context, and both left and right contexts as $Lv_1/(Lu_1 + Lu_2)$, $\ldots Lv_N/(Lu_{M-1} + Lu_M)$.
2. Prosodic (supra-segmental) parameters – F0, En_{c0}, differential F0 microintonation ($F0_{DIFF}$), jitter, shimmer, zero-crossing period, and zero-crossing frequency.

3. Basic and supplementary spectral features – first two formants (F_1, F_2), their ratio (F_1/F_2), spectral decrease (tilt), spectral centroid, spectral spread, spectral flatness, harmonics-to-noise ratio (HNR), spectral Shannon entropy (SHE).

Statistical analysis of these speech features yields various STPs: basic low-level statistics (mean, median, relative max/min, range, dispersion, standard deviation, etc.) and/or high-level statistics (flatness, skewness, kurtosis, covariance, etc.) for the subsequent evaluation process. The block diagram of creation of the speech feature databases can be seen in Fig. 2.

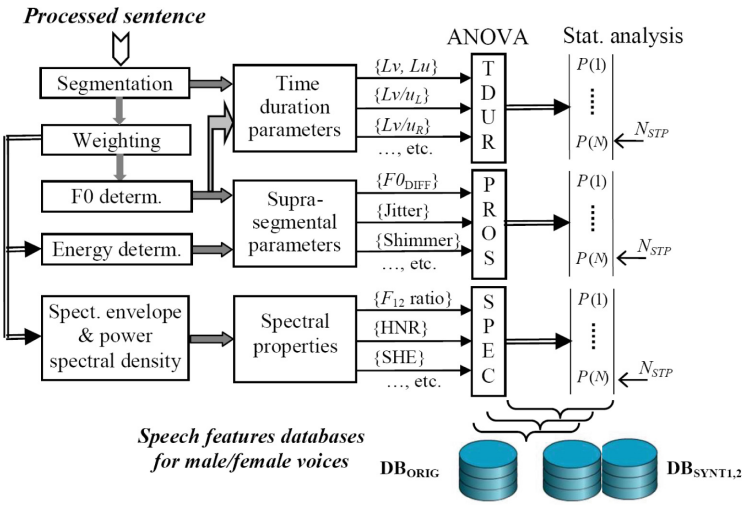

Fig. 2. Block diagram of speech feature databases creation from time durations, prosodic parameters, spectral properties, and their statistical parameters.

3 Material, Experiments and Results

The synthetic speech produced by the Czech TTS system based on the unit selection (USEL) synthesis method [12] and the sentences uttered by four professional speakers – 2 males (M1 and M2) and 2 females (F1 and F2) were used in this evaluation experiment. The main speech corpus was divided into three subsets: the first one consists of the original speech uttered by real speakers (further called as *Orig*), the second and third ones comprise synthesized speech signals produced by the TTS system with voices based on the corresponding original speaker using two different synthesis methods: with a rule-based prosody manipulation (TTSbase – *Synt1*) [13] and a modified version of the USEL method that reflects the final syllable status (TTSsyl – *Synt2*) [14]. The collected database consists of 50 sentences from each of four original speakers (200 in total), next sentences of two synthesis types giving 50 + 50 sentences from the male voice

M1 and 40 + 40 ones from the remaining speakers M2, F1, and F2. Speech signals of declarative and question sentences were sampled at 16 kHz and their duration was from 2.5 to 5 s. The main orientation of the performed experiments was to test functionality of the developed automatic evaluation system in every functional block of Fig. 1 – calculated histograms and statistical parameters are shown in demonstration examples in Figs. 3, 4 and 5. Three auxiliary comparison experiments were realized, too, with the aims to analyse:

1. effect of the number of used statistical parameters $N_{STP} = \{3, 5, 7, 10\}$ on the obtained evaluation results – see numerical comparison of values in Table 1 for the speakers M1 and F1,
2. influence of the used type of speech features (spectral, prosodic, time duration) on the accuracy and stability of the final evaluation results – see numerical results for speakers M1 and F1 in Table 2,
3. impact of the number of analysed speech signal frames on the accuracy and stability of the evaluation process – compare values for limited (15 + 15 + 15 sentences for every speaker), basic (25 + 25 + 25 sentences), and extended (50 + 40 + 40) testing sets in Table 3 for the speakers M1 and F1.

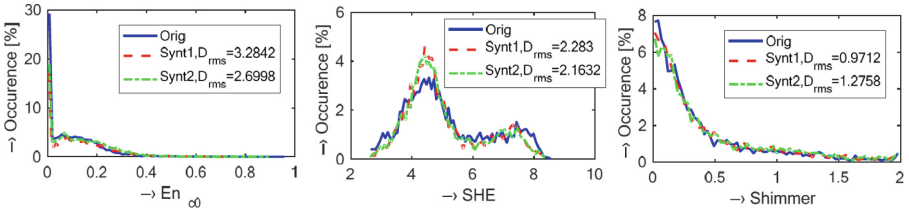

Fig. 3. Histograms of spectral and prosodic features En_{c0}, SHE, Shimmer together with calculated RMS distances between the original and the respective synthesis for the male speaker M1, using the basic testing set of 25 + 25 + 25 sentences.

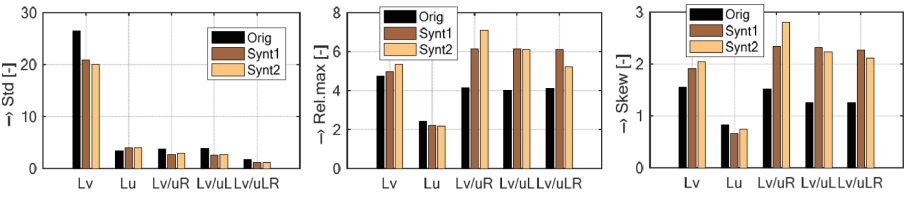

Fig. 4. Comparison of selected statistical parameters std, relative maximum, skewness calculated from values of five basic TDUR features, for the female speaker F1 and the basic testing set.

Finally, numerical comparison with the results obtained by the listening test was performed using the extended testing set. The maximum score using the

determined STPs and the mixed feature types (spectral + prosodic + time duration) is evaluated for each of four speakers – see the values in Table 4.

Subjective quality of the same utterance generated by two different approaches to prosody manipulation in the same TTS synthesis system (TTSbase and TTSsyl) was evaluated by a preference listening test. Four different male and female voices were used, each to synthesize 25 pairs of randomly selected utterances, so that the whole testing set was made up of 100 sentences. The order of two synthesized versions of the same utterance was randomized too, to avoid bias in evaluation by recognition of the synthesis method. Twenty two evaluators (8 women and 14 men) within the age range from 20 to 55 years of age participated in the listening test experiment open from 7th to 20th March 2017. The listeners were allowed to play the audio stimuli as many times as they

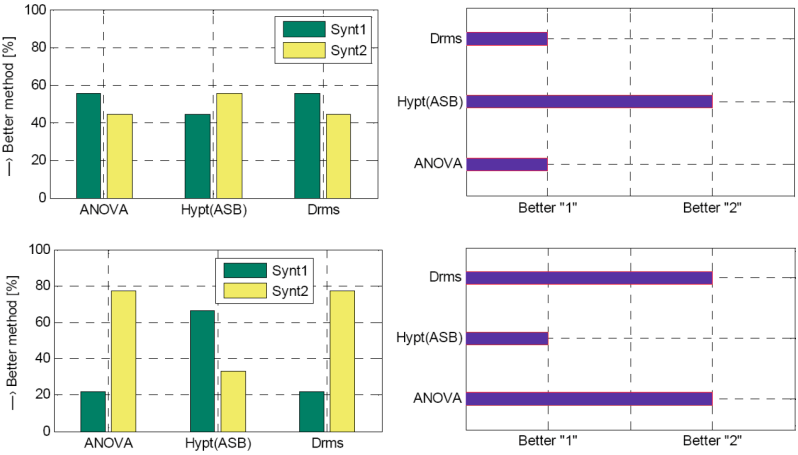

Fig. 5. Visualization of partial percentage results per three evaluation methods together with final decisions for speakers M1 (upper set of graphs) and F1 (bottom set), using only basic spectral properties from the basic set of sentences, $N_{STP} = 3$.

Table 1. Influence of the number of used statistical parameters on partial evaluation results for speakers M1 and F1, when spectral properties and prosodic parameters are used.

$N_{\mathrm{STP}}[-]^{(\mathrm{A})}$	Male speaker M1		Female speaker Fl	
	Partial	Final$^{(\mathrm{B})}$	Partial	Final$^{(\mathrm{B})}$
3	1 (65%), 2 (35%)	"**1**"	1 (60%), 2 (40%)	"**2**"
5	1 (67%), 2 (33%)	"**1**"	1 (48%), 2 (52%)	"**0**"
7	1 (71%), 2 (29%)	"**1**"	1 (44%), 2 (56%)	"**1**"
10	1 (73%), 2 (27%)	"**1**"	1 (37%), 2 (63%)	"**1**"

$^{(A)}$ used basic testing set (of 25+25+25 processed sentences),
$^{(B)}$ used "1"= TTSbase better, "0"= Similar, "2"= TTSsyl better.

Table 2. Influence of the used type of speech features (spectral, prosodic, time duration) on the accuracy and stability of the evaluation results for speakers M1 and F1.

Speech feature types[A]	Male speaker M1		Female speaker F1	
	Partial	Final[B]	Partial	Final[B]
Spectral only	1 (63%), 2 (37%)	"**1**"	1 (54%), 2 (46%)	'**1**"
Spectral+prosodic	1 (58%), 2 (42%)	"**1**"	1 (52%), 2 (48%)	"**0**"
Spectral+prosodic+ time duration	1 (46%), 2 (54%)	"**2**"	1 (44%), 1 (56%)	"**2**"

[A] used basic testing set (of 25+25+25 processed sentences), the maximum of determined STPs is applied.
[B] used "1"= TTSbase better, "0"= Similar, "2"= TTSsyl better.

Table 3. Partial evaluation results for different lengths of used speech databases for speakers M1 and F1 using only time duration features.

Speech corpus (No of sentences)[A]	Male speaker M1		Female speaker Fl	
	Partial	Final[B]	Partial	Final[B]
Limited (15+15+15)	1 (36%), 2 (64%)	"**2**"	1 (49%), 2 (51%)	"**0**"
Basic (25+25+25)	1 (29%), 2(71%)	"**2**"	1 (44%), 2 (56%)	"**2**"
Extended (50+40+40)	1 (22%), 2 (78%)	"**2**"	1 (37%), 1 (63%)	"**2**"

[A] per type of Orig+Syntl+ Synt2, the maximum of determined STPs is applied.
[B] used "1"= TTSbase better, "0"= Similar, "2"= TTSsyl better.

Table 4. Final comparison of objective and subjective evaluations for all four speakers.

Speaker	Automatic evaluation[A]		Listening test[B]		
	Partial	Final	"**1**"	"**0**"	"**2**"
Ml (AJ)	1 (40.7%), 2 (59.3%)	"**2**"	21.3%	20.0%	58.7%
M2 (JS)	1 (44.9%), 2 (55.1%)	"**2**"	16.5%	27.1%	56.4%
Fl (KI)	1 (44.4%), 2 (55.6%)	"**2**"	13.1%	21.8%	53.6%
F2 (SK)	1 (46.1%), 2 (54.9%)	"**2**"	17.1%	29.3%	58.5%

[A] used extended set of processed sentences, the maximum of determined STPs and all three types of speech features are applied.
[B] used evaluation as "1"= TTSbase better, "0"= Similar, "2"= TTSsyl better.

wished; low acoustic noise conditions and headphones were advised. Playing of the stimuli was followed by the choice between "*A sounds better*", "*A sounds similar to B*", or "*B sounds better*" [14]. The results obtained in this way were further compared with the objective results of the currently proposed system of automatic evaluation.

4 Discussion and Conclusion

The performed experiments have confirmed that the proposed evaluation system is functional and produces results comparable with the standard listening test method as documented by numerical values in Table 4. Basic analysis of the obtained results shows principal importance of application of all three types of speech features (spectral, supra-segmental, time-duration) for complex evaluation of synthetic speech. This is relevant especially when the compared synthesized speech signals differ only in their prosodic manipulation, as in the case of this speech corpus. Using only the spectral features brings non-stable or contradictory results, as shown in "Final" columns of Table 2. The detailed analysis showed principal dependence of the correctness of evaluation on the number of used statistical parameters – compare particularly the values for the female voice in Table 1. For $N_{STP} = 3$ the second synthesis type was evaluated as better and increase of the number of parameters to 5 resulted in considering both methods as similar. Further increase of the number of parameters to 7 and 10 gave stable results with preference of the first synthesis type. Additional analysis has shown that a minimum number of speech frames must be processed to achieve correct statistical evaluation and significant statistical differences between the original and tested STPs derived from the same speaker. If these were not fulfilled, the final decision of the whole evaluation system would not be stable and no useful information would be got by "0"category of the automatic evaluation system equivalent to "*A sounds similar to B*" in the subjective listening test. Tables 1, 2 and 3 show this effect for the female speaker F1. In general, the tested evaluation system detects and classifies male speakers better than female ones. It may be caused by higher variability of female voices and its effect to the supra-segmental area (changes of energy and F0), the spectral domain, and the changes in time duration relations.

In the near future, we will try to collect larger speech databases, including greater number of speakers. Next, in the databases, there will be incorporated more different methods of speech synthesis (HMM, PSOLA, etc.) produced by more TTS systems in other languages – English, German, etc. In this way, we will carry out complex testing of automatic evaluation with the final aim to substitute subjective evaluation based on the listening test method.

References

1. Grůber, M., Matoušek, J.: Listening-test-based annotation of communicative functions for expressive speech synthesis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS (LNAI), vol. 6231, pp. 283–290. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15760-8_36
2. Monte-Moreno, E., Chetouani, M., Faundez-Zanuy, M., Sole-Casals, J.: Maximum likelihood linear programming data fusion for speaker recognition. Speech Commun. **51**(9), 820–830 (2009)
3. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. **3**, 72–83 (1995)

4. Xu, L., Yang, Z.: Speaker identification based on state space model. Int. J. Speech Technol. **19**(2), 407–414 (2016)
5. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E., Torres-Carrasquillo, P.A.: Support vector machines for speaker and language recognition. Comput. Speech Lang. **20**(2–3), 210–229 (2006)
6. Lee, C.Y., Lee, Z.J.: A novel algorithm applied to classify unbalanced data. Appl. Soft Comput. **12**, 2481–2485 (2012)
7. Mizushima, T.: Multisample tests for scale based on kernel density estimation. Stat. Probab. Lett. **49**, 81–91 (2000)
8. Hussain, T., Siniscalchi, S.M., Lee, C.C., Wang, S.S., Tsao, Y., Liao, W.H.: Experimental study on extreme learning machine applications for speech enhancement. IEEE Accesss **5**, 25542 (2017)
9. van Santen, J.P.H.: Segmental duration and speech timing. In: Sagisaka, Y., Campbell, N., Higuchi, N. (eds.) Computing Prosody. Springer, New York (1997). https://doi.org/10.1007/978-1-4612-2258-3_15
10. Martinez, C.C., Cassol, M.: Measurement of voice quality, anxiety and depression symptoms after therapy. J. Voice **29**(4), 446–449 (2015)
11. Rietveld, T., van Hout, R.: The t test and beyond: recommendations for testing the central tendencies of two independent samples in research on speech, language and hiering pathology. J. Commun. Disord. **58**, 158–168 (2015)
12. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Atlanta (Georgia, USA), pp. 373–376 (1996)
13. Tihelka, D., Kala, J., Matoušek, J.: Enhancements of Viterbi search for fast unit selection synthesis. In: Proceedings of INTERSPEECH 2010, Makuhari, Japan, pp. 174–177 (2010)
14. Jůzová, M., Tihelka, D., Skarnitzl, R.: Last syllable unit penalization in unit selection TTS. In: Ekštein, K., Matoušek, V. (eds.) TSD 2017. LNCS (LNAI), vol. 10415, pp. 317–325. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64206-2_36