Journal of ELECTRICAL ENGINEERING, VOL 71 (2020), NO2, 78–86

sciendo

PAPER

Automatic statistical evaluation of quality of unit selection speech synthesis with different prosody manipulations

Jiří Přibil[1,2], Anna Přibilová[1], Jindřich Matoušek[2]

Quality of speech synthesis is a crucial issue in comparison of various text-to-speech (TTS) systems. We proposed a system for automatic evaluation of speech quality by statistical analysis of temporal features (time duration, phrasing, and time structuring of an analysed sentence) together with standard spectral and prosodic features. This system was successfully tested on sentences produced by a unit selection speech synthesizer with a male as well as a female voice using two different approaches to prosody manipulation. Experiments have shown that for correct, sharp, and stable results all three types of speech features (spectral, prosodic, and temporal) are necessary. Furthermore, the number of used statistical parameters has a significant impact on the correctness and precision of the evaluated results. It was also demonstrated that the stability of the whole evaluation process is improved by enlarging the used speech material. Finally, the functionality of the proposed system was verified by comparison of the results with those of the standard listening test.

Keywords: listening test, objective and subjective evaluation, quality of synthetic speech, statistical analysis

1 Introduction

Speech quality can be evaluated by various subjective and objective measures and methods. Subjective assessment is usually based on the perception of intelligibility, naturalness, similarity, quality, *etc*. The most used subjective measures are the mean opinion score (with application in the area of emotional speech recognition [1], speech corpus annotation [2], *etc*), the comparison category rating [3], the preference test [4] (ABX test representing a choice from two alternatives – *eg* to find the best resemblance between the original and the target synthetic speech), or the choice from among various application-specific possibilities. In the objective estimation, the speech spectrum may be compared using various methods, or pitch and voicing errors may be evaluated. Most objective evaluation methods use various types of spectral or prosodic features. The most used ones are mel frequency cepstral coefficients (MFCC) with a large application area in sound classification tasks [5], automatic speech or speaker recognition [6], age estimation, classification of expressive speech, *etc*. The MFCCs are subsequently used in the evaluation process based on a statistical analysis of variance (ANOVA), hypothesis tests, *etc* [7]. The automatic speech recognition (ASR) approaches [8] or the ASR based on hidden Markov models [9] are often used for evaluation of the synthetic speech. Perceptual evaluation of speech quality (PESQ; ITU-T recommendation P.862) [10]) was the most popularly used objective measure incorporating a perceptual model for speech quality assessment for telephony applications and narrow-band speech coders [11]. Both approaches (subjective and objective ones) are also combined [12, 13]. While in the synthetized speech by coders for telephone applications the main important parameters are the bandwidth (*eg* narrow-band, wideband, super-wideband), the sampling frequency, the bitrate, *etc* determining the final quality of the synthetized speech, in our case of the text-to-speech (TTS) synthesis they are irrelevant.

The main motivation of this work was to design, realize, and test a system for automatic evaluation of speech quality as an alternative to the standard subjective listening test. Previous analysis has shown that suprasegmental features derived from time durations of voiced and unvoiced speech parts [14] must be comprised in the complex automatic system evaluating the quality of synthetic speech by comparison of two or more utterances synthesized by different TTS systems. Speech features based on MFCCs or other spectral properties cannot render changes in the time structure caused by prosody manipulation during the process of speech synthesis. Therefore, time-domain speech features are also necessary for comparison of a synthetic utterance generated by a TTS system and an original speech of a given speaker – differing in the way of phrase creation, the speed of the utterance, and/or the time-domain changes in prosody production, *etc*. This article describes the function of the proposed system for automatic assessment of synthetic speech signal quality in terms of similarity with the original by evaluation of features derived from time durations of voiced and unvoiced speech parts. The proposed system for automatic objective evaluation could replace a subjective method of listening tests when it is difficult to discern audible differences or there is a problem with re-

[1] Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia, {jiri.pribil. anna.pribilova}@savba.sk,
[2] Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic, jmatouse@kky.zcu.cz

DOI:10.2478/jee-2020–0012, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

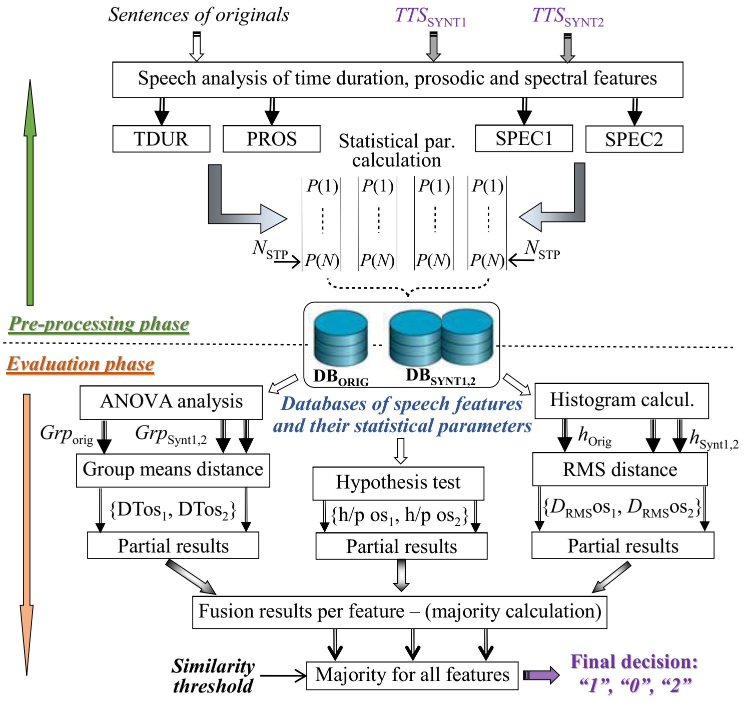
Sentences of originals
TTS_{SYNT1}
TTS_{SYNT2}
Speech analysis of time duration, prosodic and spectral features
TDUR
PROS
Statistical par. calculation
SPEC1
SPEC2
$P(1)$
$P(N)$
N_{STP}
Pre-processing phase
Evaluation phase
$\mathrm{DB}_{\mathrm{ORIG}}$
$\mathrm{DB}_{\mathrm{SYNT1,2}}$
Databases of speech features and their statistical parameters
ANOVA analysis
Grp_{orig}
$Grp_{\mathrm{Synt1,2}}$
Group means distance
$\{\mathrm{DTos}_1, \mathrm{DTos}_2\}$
Partial results
Hypothesis test
$\{\mathrm{h/p\ os}_1, \mathrm{h/p\ os}_2\}$
Partial results
Histogram calcul.
h_{Orig}
$h_{\mathrm{Synt1,2}}$
RMS distance
$\{D_{\mathrm{RMS}}\mathrm{os}_1, D_{\mathrm{RMS}}\mathrm{os}_2\}$
Partial results
Fusion results per feature – (majority calculation)
Similarity threshold
Majority for all features
Final decision: "1", "0", "2"

Fig. 1. Block diagram of the automatic system for evaluation of the synthetic speech, including the speech database pre-processing

production for a greater number of listeners in the same auditory conditions, *etc*.

Then, the contribution deals at length with a procedure of acquisition of voiced and unvoiced time durations in the speech signal by the detected fundamental frequency and energy curves together with a method for determination of standard spectral and prosodic features. The temporal features themselves are computed using different statistical parameters for the creation of a database classified by individual male/female speakers. Next, it is focused on the description of experiments verifying functionality and stability of evaluation of the synthetic speech signal quality by the proposed automatic system. Finally, the results are compared with those of the listening tests using the same synthetic speech corpus.

2 USED EVALUATION METHOD

2.1 Basic principles of the applied method of speech quality evaluation

The whole automatic evaluation process starts with the initial phase of creating the databases from the analysed male and female natural utterances and the synthetic ones generated by different methods of TTS synthesis, different synthesis parameters, *etc*. Each of these basic databases consists of the time duration features (TDUR), the supra-segmental prosodic parameters (PROS), and the basic and supplementary spectral properties (SPEC1, SPEC2). In the next step, separate calculations of the statistical parameters (STP) are made for every speaker and every speech feature. The determined statistical parameters together with the speech feature values are stored for next use in separate databases depending on the used input signal ($\mathrm{DB}_{\mathrm{ORIG}}$, $\mathrm{DB}_{\mathrm{SYNT1}}$, $\mathrm{DB}_{\mathrm{SYNT2}}$) and the speaker (male/female) - see the upper part of the block diagram in Fig. 1. Statistical analysis of the speech features saved in these databases yields various STP: basic low-level statistics (mean, median, relative max/min, range, dispersion, standard deviation, *etc*) and/or high-level statistics (flatness, skewness, kurtosis, covariance, *etc*).

The second phase is represented by practical evaluation of the processed data: construction of histograms of feature value distribution, calculation of the ANOVA statistics and probability assessment of the hypothesis resulting from the Ansari-Bradley test (ASB). It makes a decision about equality of two distributions or difference in their variances. A similar test is the Wilcoxon test (Mann-Whitney U test) determining equality of two distributions by their medians or inequality of the medians [15, 16]. The output of these tests is also the probability of the null hypothesis about identical distributions. If this probability is higher than a significance level, the hypothesis logical value is zero and the null hypothesis cannot be rejected. Otherwise, the logical value is one and the null hypothesis can be rejected.

Three output parameters for comparison of the values between the original speech *Orig* and the synthesized one *Synt*1/*Synt*2 are further calculated: root-mean-square (RMS) distances (Dh_{RMS}) between the preliminary calculated histograms for each of speech features, Da_{GRP} – distances between group means of the ANOVA, and

Table 1. Used types of speech features

Feature type	Feature description
Prosodic (PROS)	F_0, signal energy (En_{c_0}), differential F_0 $(F_{0\text{DIFF}})$, jitter, shimmer, zero-crossing period, zero-crossing frequency
Basic spectral (SPEC1)	first two formants (F_1, F_2), their ratio (F_1/F_2), spectral tilt (decrease), spectral spread, first four cepstral coefficients $(c_1\text{–}c_4)$
Supplementary spectral (SPEC2)	harmonics-to-noise ratio, spectral centroid, spectral flatness, Shannon, Rényi, and Tsallis spectral entropies

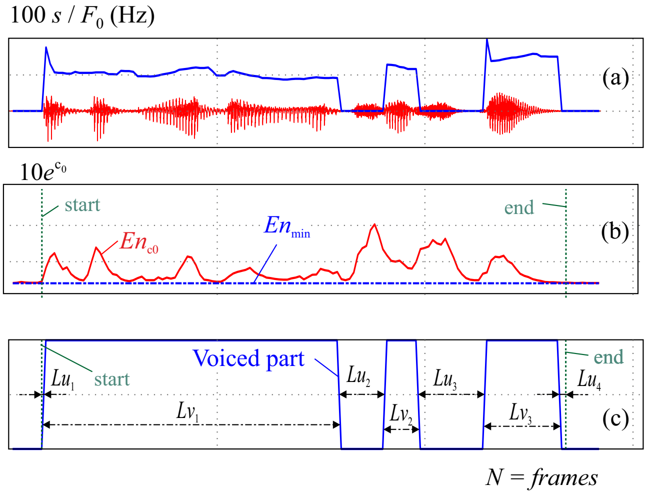

Fig. 2. Example of voicing determination from F_0 contour with applied energy threshold: (a) – speech signal together with F_0 contour, (b) – En_{c0} contour, applied threshold $En_{\min} = 0.02$, eliminated parts at the beginning and at the end, (c) – finally determined voiced/unvoiced parts

the hypothesis value with its probability (h/p). For every speech feature, the obtained Dh_{RMS} values and their STP are next used to compare similarity to the original - see the evaluation phase in the bottom part of the block diagram in Fig. 1. The partial decision is determined for the total number of N_{STP} processed values by applying the majority function to each of the obtained partial results. Then, the majority function is applied again to these partial decision values to get the final decision about the proximity of the tested synthetic speech produced by the TTS system to the original speech utterance. The output value "1" ("2") means *Synt*1 (*Synt*2) close to *Orig* and "0" denotes similarity due to differences below the set threshold. This objective evaluation result corresponds to the subjective listening test choice "A sounds similar to B" [2] with small or indiscernible perceptual differences. The final decision about better synthesis is determined using the majority function of the partial results.

2.2 Determination of various types of speech features

Before the creation of databases of speech features and STP, the speech signal is processed in weighted frames. The signal energy calculated from the first cepstral coefficient c_0 (En_{c_0}) is used to eliminate speech pauses at the beginning and at the end of the uttered sentence. Only voiced or unvoiced frames with the energy higher than the chosen threshold $En_{\min}$ are used in the next processing - see a demonstration example in Fig. 2(a),(b). The analysed signal should always begin and end with unvoiced parts, but the energy lower than a threshold may cut these parts. Consequently, if the speech signal begins and/or ends with a voiced part, the unvoiced part with the mean duration of all unvoiced parts is inserted at the beginning and/or the end of this signal. Thus, the F_0 contour of the analysed sentence can be divided into N voiced parts and $N+1$ unvoiced parts with various durations as documented in Fig. 2(c). In this way, the following five types of TDUR features are determined:

1. Lv – absolute duration of a voiced part in frames (N values),
2. Lu – absolute duration of an unvoiced part in frames ($N+1$ values),
3. $L_{\text{V/U-L}}$ – a ratio of absolute durations of voiced and unvoiced parts adjacent to the left of the former: $Lv_1/Lu_1, \ldots, Lv_N/Lu_N$,
4. $L_{\text{V/U-R}}$ – a ratio of absolute durations of voiced and unvoiced parts adjacent to the right of the former: $Lv_1/Lu_2, \ldots, Lv_N/Lu_{N+1}$,
5. $L_{\text{V/U-LR}}$ – a ratio of the duration of a voiced part and the mean duration of unvoiced parts adjacent to the left and right: $Lv_1/(Lu_1 + Lu_2), \ldots, Lv_N/(Lu_N + Lu_{N+1})$.

Apart from the TDUR features, the contours of F_0 and signal energy are used to determine standard suprasegmental (prosodic) parameters. Other speech features are the segmental (spectral) ones designated in each frame of the input sentence from the smoothed spectral envelope or the power spectral density, see Tab. 1.

3 MATERIAL, EXPERIMENTS AND RESULTS

3.1 Used speech material

The first of the used speech databases represents the original natural speech (further called *Orig*) consisting of declarative sentences uttered by four professional speakers – 2 males (M1 and M2) and 2 females (F1 and F2) in the Czech language. The second and third databases comprise sentences with the same contents produced by

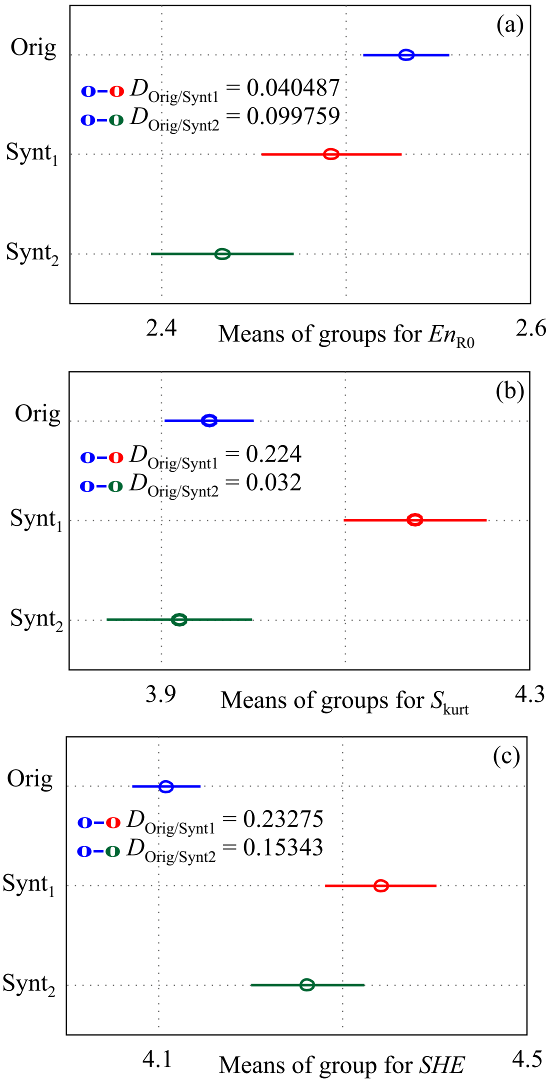

Fig. 3. Visualization of multiple comparison of group means of ANOVA applied to the speech features En$_{R0}$, S_{kurt}, SHE with calculated distances $D_{\text{Orig/Synt1,2}}$ for the male speaker M1

two different speech synthesis methods with the voices based on the original speakers of the first database. The TTS synthesizer uses either the unit selection method (USEL) [17] with the rule-based prosody manipulation (further called TTS$_{Synt1}$) [18] or the modified version of the unit selection method reflecting the final syllable status (further called TTS$_{Synt2}$) [19]. The collected database consists of 50 sentences from each of the four original speakers (200 in total) and the two types of synthesized sentences (50+50 from the male voice M1, 40+40 from the remaining speakers M2, F1, and F2). Speech signals of all processed sentences (original as well as synthetic ones) were sampled at 16 kHz and their duration ranged from 2.5 to 5 seconds. The frame length for spectral analysis depends on the mean pitch period of the speech signal. In these experiments, 24 ms frames were chosen for the male voices and 20 ms frames for the female ones so that the frame duration was at least twice the pitch period (see the speakers mean F_0 values in Tab. 2). Calculation of TDUR features was supplemented with the determination of the fundamental frequency F_0 by the autocorrelation analysis method with experimentally chosen pitch ranges from 55 Hz to 250 Hz for the male voices and from 105 Hz to 350 Hz for the female ones.

Table 2. Detailed description of used speech material in performed experiments

Speaker type	$F_{0\text{Mean}}$ (Hz)	No of sentences/T_{DUR} (s)/No of frames* Originals	TTS$_{\text{Synt1}}$	TTS$_{\text{Synt2}}$
M1 (AJ)	120	50/133/9166	50/122/9631	50/120/9561
M2 (JS)	100	50/132/9204	40/103/7735	40/100/7648
F1 (KI)	215	50/137/9596	40/102/8115	40/98/7838
F2 (SK)	195	50/141/9876	40/97/8533	40/94/7518

* frames processed excluding beginning and ending parts with low energy

Table 3. Values of the Ansari-Bradley hypothesis test for three speech features corresponding to Fig. 3.

Feature type	Hypothesis/probability values* Orig vs. Synt1	Orig vs. Synt2	Partial results
En$_{R0}$	0/0.21	1 / 0.014	Better Synt1
S_{kurt}	1/4.39 10^{-5}	0/0.23	Better Synt2
SHE	1/1.18 10^{-5}	1/4.71 10^{-5}	Similar

* for 5 % significance level

3.2 Description of performed experiments and obtained results

The main purpose of the performed experiments was to test the functionality of the designed automatic evaluation system. Partial results documenting the process of computation and comparison for each of the functional blocks are presented in a graphical as well as a numerical form. Graphs in Fig. 3 visualize multiple comparisons of group means of the ANOVA applied on the speech signal energy determined by the first autocorrelation coefficient R_0 (En$_{R0}$), the spectral kurtosis (S_{kurt}), and the Shannon spectral entropy (SHE) including calculated distances between the original and each of two tested syntheses. Table 3 shows the corresponding values of null hypothesis/probability values for 5 % significance level of the ASB hypothesis test together with partial decision results. Figure 4 represents histograms of voiced/unvoiced time duration parts and their ratios together with calculated RMS distances. Bar-graphs of selected statistical parameters of three TDUR features are shown in Fig. 5. Figure 6 presents partial percentages obtained by ANOVA, ASB hypothesis test, RMS distances from histograms of spectral and prosodic features used for the finally calculated majority result after fusion. Two auxiliary comparison experiments were realized with the aim to analyse:

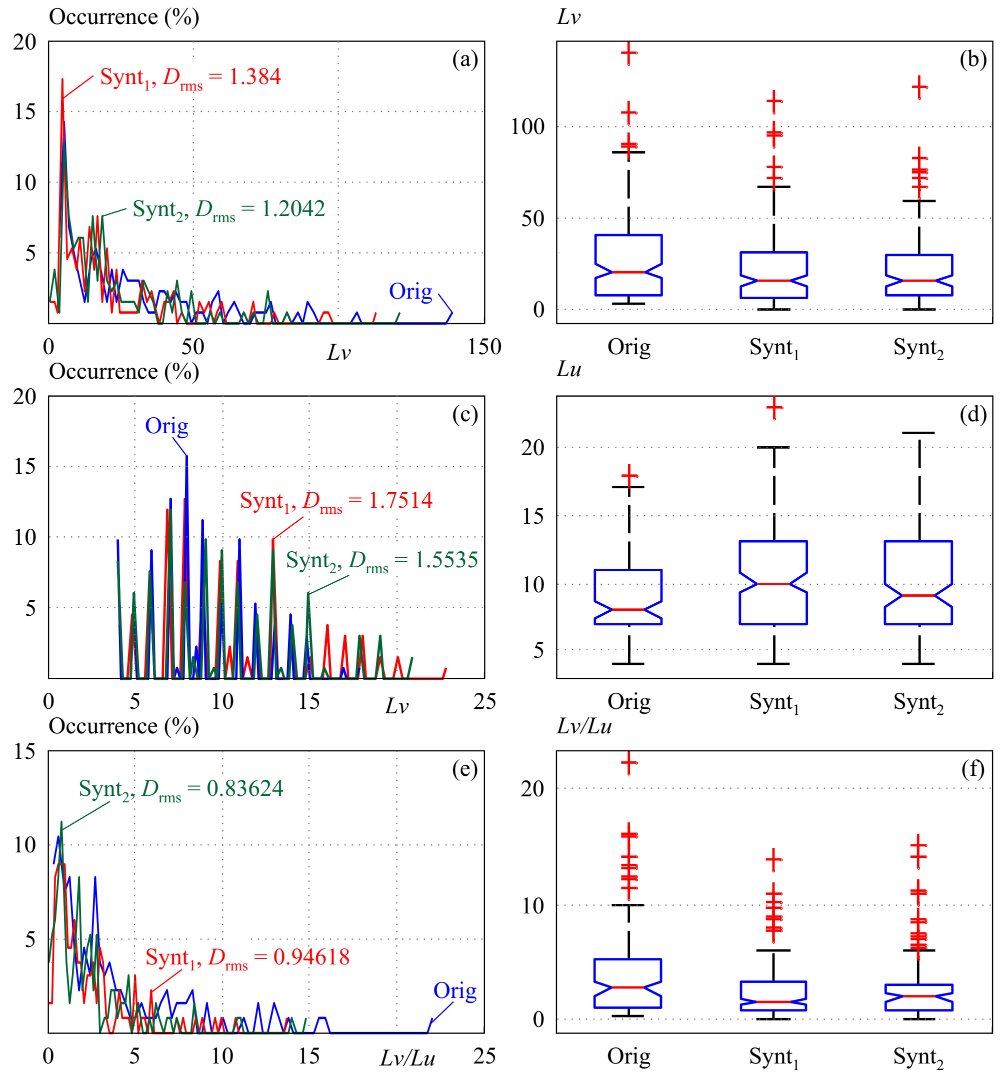

Fig. 4. Histograms for: (a) – voiced, (b) – unvoiced time duration parts, and (c) – their ratios, with calculated RMS distances between the original and the respective synthesis (left graphs), box-plot of basic statistical parameters (right graphs) for the female speaker F1

1. effect of the number of used statistical parameters $N_{\mathrm{STP}} = \{3, 5, 7, 10, 16\}$ on the obtained evaluation results, see the graphical comparison for the male speaker M1 and the female one F1 in Fig. 7,
2. influence of different types of used speech features (temporal, spectral, prosodic) on the accuracy and stability of the final evaluation results, see the numerical results for M1 and F1 in Tabs. 4 and 5.

Finally, numerical comparison with the results obtained by the listening tests for each of four tested speakers (M1,2+F1,2) was performed – see the bar-graph comparison in Fig. 8.

Subjective quality of synthetic speech was assessed by a large preference listening test. The used two variants of the TTS synthesis of a sentence were the same as in the automatic system. The final set of 100 pairs of randomly selected sentences comprised 4 different male and female voices with 25 pairs per voice. The subjective test experiment was attended by 22 listeners (14 males and 8 females) in the age from 20 to 55 years during the time period from 7th to 20th March 2017. The listeners were recommended to use headphones and to perform the test in quiet noise conditions. There was a possibility of repeated listening of every audio stimulus and then one of the following choices had to be selected “A sounds better”, “A sounds similar to B”, or “B sounds better”. The performed listening test was described in more detail in [19].

4 DISCUSSION AND CONCLUSION

The performed experiments have confirmed that the proposed automatic evaluation system is functional, and the obtained results are comparable with the ones of the standard listening test method. This fact is documented by graphical comparison in Fig. 8. It means that the basic motivation task was fulfilled in principle - the substitution of the subjective evaluation by the objective one to eliminate the main disadvantage of human assessment:

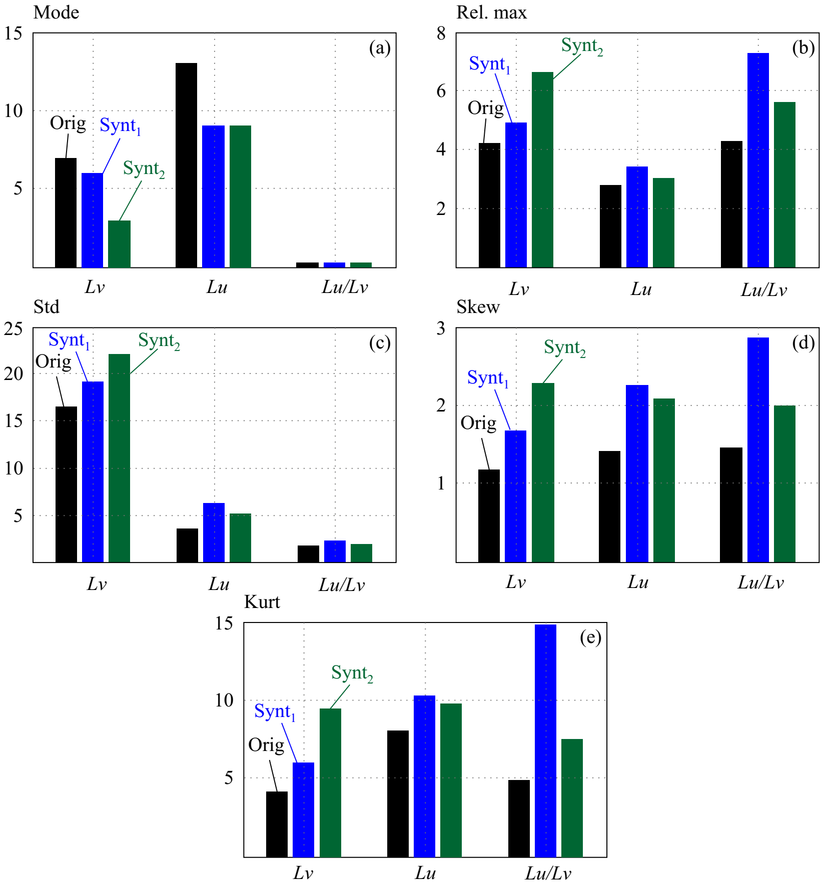

Fig. 5. Bar-graph comparison of selected statistical parameters mode, rel. max, std, kurtosis, and skewness calculated from three basic TDFs derived from the values presented in Fig. 4

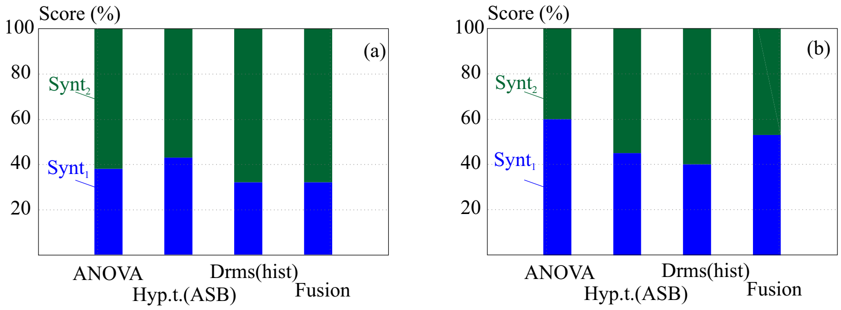

Fig. 6. Visualization of partial scores and a final score after fusion using only spectral and prosodic features for: (a) – male voice M1 (fusion of 68 % for "Better Synt2"), (b) – female voice F1 (fusion of 53 % as "Similar"); applied $N_{\mathrm{STP}} = 5$

Table 4. Influence of used types of speech features on evaluation results for the male speaker M1

Feature type	Evaluation results	
	Partial*	Final**
TDUR	2 (57 %)	Better Synt2
SPEC1+SPEC2	1 (63 %)	Better Synt1
PROS+SPEC1,2	1 (57 %)	Better Synt1
TDUR+PROS+SPEC1,2	2 (59 %)	Better Synt2

* $N_{\mathrm{STP}} = 16$ was applied, ** For 5 % similarity threshold

subjectivity, lack of reproducibility, dependence on environmental conditions, and very high time consumption.

According to a detailed analysis, the evaluation correctness depends principally on the used number of statistical parameters. It is significant mainly in the case of the testing sentences of a female voice – as documented by the bar-graph comparison in Fig. 7. Using $N_{\mathrm{STP}} = 3$, the sentences produced by the first synthesis method [19] were wrongly evaluated as better, for $N_{\mathrm{STP}} = 5$, the decision falls to the "Similar" category, and only for $N_{\mathrm{STP}} \geq 7$ the obtained results are correct ($Synt2$ is better) and stable. Next auxiliary analysis shows principal importance of application of all types of speech features (temporal, prosodic, and spectral) for correct complex evaluation of the synthetic speech. This is relevant especially for

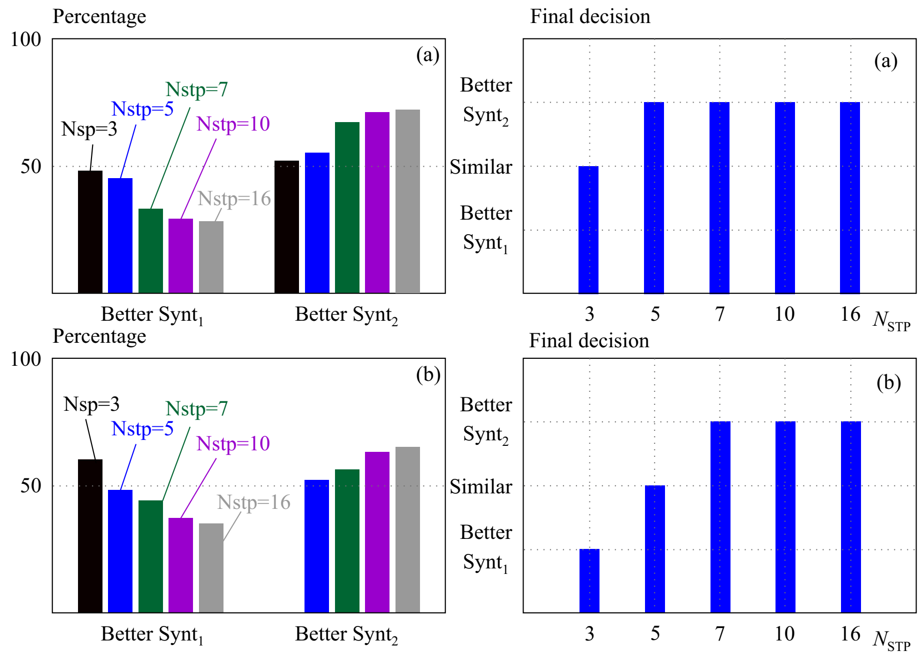

Fig. 7. Influence of the number of used statistical parameters on partial evaluation results together with visualization of the final evaluation decision for: (a) – male voice M1, (b) – female voice F1

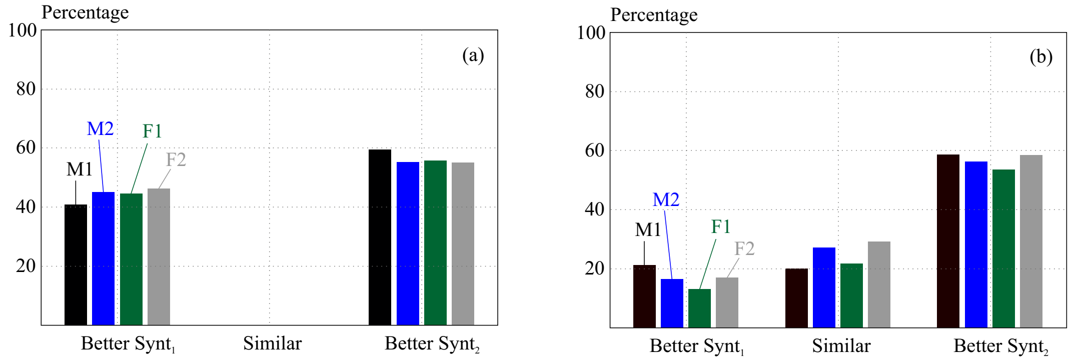

Fig. 8. Final comparison of objective and subjective evaluations for all four speakers: (a) – results of the automatic evaluation for used mixed speech features and maximum N_{STP}, (b) – listening test results [19]

Table 5. Influence of used types of speech features on evaluation results for the female F1.

Feature type	Evaluation results	
	Partial*	Final**
TDUR	2 (56 %)	Better Synt2
SPEC1+SPEC2	1 (58 %)	Better Synt1
PROS+SPEC1,2	1 (52 %)	Similar
TDUR+PROS+SPEC1,2	2 (57 %)	Better Synt2

* $N_{\mathrm{STP}} = 16$ was applied, ** For 5 % similarity threshold

the compared synthesized speech signals differing only in prosodic manipulation in this speech corpus. Using only the spectral features brings non-stable or contradictory results and also application of the mixed prosodic and spectral features without the time-duration ones gives no correct results – as confirmed by the obtained final results in the last columns in Tab. 4 and Tab. 5. As regards the speaker gender, the male voice is classified better than the female one by this evaluation system. It may be caused by higher variability of female voices and its effect on the supra-segmental area (changes in energy and F_0), the spectral domain, and the changes in time duration relations.

The building of the database of temporal features is rather time-consuming and must be processed off-line. In addition, the current realization of the whole automatic evaluation system was implemented in the Matlab environment. The computational complexity must be analysed and the algorithm must be optimized to increase the computing speed and to enable real-time evaluation of the statistical parameters. Once the critical points are found, the algorithm can be implemented in a higher programming language.

Soon, we will try to collect larger speech databases, including a greater number of speakers. Then, in the databases, more different methods of speech synthesis based on a deep neural network (DNN) paradigm [20], such as long short-term memory (LSTM) networks [20], WaveNet [21] or WaveRNN [22] will be incorporated. At present, we also can produce the synthesis speech in the Slovak language which is similar to Czech [23], so the application of Slovak in the proposed evaluation system is expected. Finally, we will try to improve the results of objective evaluation by adding other statistical analysis methods or parameters, such as intraclass correlation or Fleiss' kappa, by inputting them to the fusion block for final decision determination.

Acknowledgements

The work has been supported by the Czech Science Foundation GA CR, project No GA19-19324S (J. Matoušek, and J. Přibil), by the Slovak Scientific Grant Agency project VEGA 2/0003/20 (J. Přibil), and the COST Action CA16116 (A. Přibilová).

References

[1] A. Zelenik and Z. Kacic, "Multi-Resolution Feature Extraction Algorithm in Emotional Speech Recognition", *Elektronika ir Elektrotechnika*, vol. 21, no. 5, pp. 54–58, 2015, DOI: 10.5755/j01.eee.21.5.13328.

[2] M. Grůber and J. Matoušek, "Listening-Test-Based Annotation of Communicative Functions for Expressive Speech Synthesis", *P. Sojka, A. Horak, I. Kopecek, K. Pala (eds.): Text, Speech, and Dialogue* (TSD) 2010, LNCS, vol. 6231, pp. 283–290, Springer 2010.

[3] P. C. Loizou, "Speech Quality Assessment", *W. Tao, et al.(eds): Multimedia Analysis, Processing and Communications. Studies Computational Intelligence*, vol. 346, pp. 623–654, Springer, Berlin, Heidelberg, 2011, DOI:10.1007/978-3-642-19551-8_23.

[4] H. Ye and S. Young, "High Quality Voice Morphing", *ICASSP 2004 Proceedings. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 17-21 May 2004, Montreal, Canada, DOI:10.1109/ICASSP.2004.1325909.

[5] M. Adiban, B. BabaAli and S. Shehnepoor, "Statistical Feature Embedding for Heart Sound Classification", *Journal of Electrical Engineering*, vol. 70, no. 4, pp. 259–272, 2019, DOI: 10.2478/jee-2019-0056.

[6] B. Boilović, B. M. Todorović and M. Obradović, "Text-Independent Speaker Recognition using Two-Dimensional Information Entropy", *Journal of Electrical Engineering*, vol. 66, no. 3, pp. 169–173, 2015, DOI: 10.1515/jee-2015-0027.

[7] C. Y. Lee and Z. J. Lee, "A Novel Algorithm Applied to Classify Unbalanced Data", *Applied Soft Computing*, vol. 12, pp. 2481–2485, 2012, DOI: 10.1016/j.asoc.2012.03.051.

[8] R. Vích, J. Nouza and M. Vondra, "Automatic Speech Recognition Used for Intelligibility Assessment of Text-to-Speech Systems", *A. Esposito et al. (eds.): Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, LNCS, vol. 5042, pp. 136–148, Springer 2008.

[9] M. Cerňak, M. Rusko and M. Trnka, "Diagnostic Evaluation of Synthetic Speech using Speech Recognition", *Procs. of the 16th International Congress on Sound and Vibration* (ICSV16), Kraków, Poland, 5-9 July, p. 6, 2009, https://pdfs.semanticscholar.org/502b/f1d8bfb0cc90cd3defcc9d479d9a97b23b66.pdf.

[10] S. Möller, and J. Heimansberg, "Estimation of TTS Quality Telephone Environments Using a Reference-free Quality Prediction Model", *Second ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, Berlin, Germany, September 2006, pp. 56–60, ISCA Archive, http://www.isca-speech.org/archive_open/pqs2006.

[11] D.-Y. Huang, "Prediction of Perceived Sound Quality of Synthetic Speech", *Procs. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (APSIPA ASC), 2011 Xi'an, China, October 18-21, 2011, p. 6, http://www.apsipa.org/proceedings_2011/pdf/APSIPA100.pdf.

[12] S. Möller *et al*, "Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems", *2010*, INTERSPEECH-2010, pp. 1325–1328, https://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_1325.pdf.

[13] F. Hinterleitner *et al*, "Predicting the Quality of Synthesized Speech using Reference-Based Prediction Measures", *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, Session: Sprachsynthese-Evaluation und Prosodie, 2011, pp. 99–106, TUDpress, Dresden, http://www.essv.de/paper.php?id=14.

[14] J. P. H. van Santen, "Segmental Duration and Speech Timing", *Y. Sagisaka, N.Campbell, N.Higuchi (eds.): Computing Prosody*, Springer, New York, NY, pp. 225–248, 1997.

[15] C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[16] V. Rodellar-Biarge, D. Palacios-Alonso, V. Nieto-Lluis, and P. Gomez-Vilda, "Towards the search of detection speech-relevant features for stress", *Expert Systems*, vol. 32, no.6, pp. 710-718, 2015.DOI: 10.1111/exsy.12109.

[17] A. J. Hunt and A. W. Black, "Unit Selection a Concatenative Speech Synthesis System using a Large Speech Database", *Proceedings of the IEEE International Conference on Acoustics*, Speech and Signal Processing (ICASSP), Atlanta (Georgia, USA), pp. 373–376, 1996, DOI: 10.1109/ICASSP.1996.541110.

[18] J. Kala and J. Matoušek, "Very Fast Unit Selection using Viterbi Search with Zero-Concatenation-Cost Chains", *Proceedings of IEEE International Conference on Acoustics*, Speech and Signal Processing (ICASSP 2014), Florence, Italy, pp. 2569–2573, 2014.

[19] M. Jůzová, D. Tihelka and R. Skarnitzl, "Last Syllable Unit Penalization Unit Selection TTS", *K. Ekstein and V. Matousek (eds.): Text, Speech, and Dialogue* (TSD 2017), LNAI vol. 10415, pp. 317–325, 2017, DOI: 10.1007/978-3-319-64206-2 36.

[20] D. Tihelka, Z. Hanzlíček, M. Jůzová, J. Vít, J. Matoušek and M. Grůber, "Current State of Text-to-Speech System ARTIC: A Decade of Research on the Field of Speech Technologies", *P. Sojka, A.Horák, I.Kopeček, and K. Pala (eds): Text, Speech, and Dialogue* (TSD 2018), LNAI 11107, pp. 369–378, 2018, DOI: doi.org/10.1007/978-3-030-00794-2_40.

[21] Z. Hanzlíček, J. Vít, and D. Tihelka, "WaveNet-Based Speech Synthesis Applied to Czech – A Comparison with the Traditional Synthesis Methods", *P. Sojka, A.Horák, I.Kopeček, and K. Pala (eds): Text, Speech, and Dialogue* (TSD 2018), LNAI 11107, pp. 445–452, 2018, DOI: 10.1007/978-3-030-00794-2_48.

[22] J. Vít, Z. Hanzlíček and J. Matoušek, "Czech Speech Synthesis with Generative Neural Vocoder", *K. Ekštein (ed.): Text, Speech, and Dialogue* (TSD 2019), LNAI 11697, pp. 307–315, 2019, DOI: 10.1007/978-3-030-27947-9_26.

[23] J. Matoušek, D. Tihelka and J. Psutka, "New Slovak Unit-Selection Speech Synthesis ARTIC TTS System", *Proceedings of the International Multiconference of Engineers and Computer Scientists* (IMECS), San Francisco, USA, 2011.

Received 1 October 2019

Jiří Přibil was born in 1962 in Prague, Czechoslovakia. He received his MSc degree in computer engineering in 1991 and his PhD degree in applied electronics in 1998 from the Czech Technical University in Prague. At present, he is a senior scientist at the Department of Imaging Methods in the Institute of Measurement Science, Slovak Academy of Sciences, Bratislava.

His research interests are signal and image processing, speech analysis and synthesis, and text-to-speech systems.

Anna Přibilová received her MSc and PhD degrees from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology (FEEIT SUT) in 1985 and 2002, respectively. In 2014 she has become an associate professor at the Institute of Electronics and Photonics of the FEEIT SUT in Bratislava. At present, she is a scientist at the Department of Biomeasurements in the Institute of Measurement Science, Slovak Academy of Sciences, Bratislava. Her research lies in the area of biomedical signal measurement, processing, and analysis.

Jindřich Matoušek received his MSc and PhD degrees from the Faculty of Applied Sciences (FAS), University of West Bohemia (UWB), Pilsen, Czech Republic in 1997 and 2001, respectively. Since 1999 he has been working as a researcher at the Department of Cybernetics FAS UWB, and since 2012 he also has been working as a member of a research team of the New Technology for Information Society (NTIS) centre at UWB. In 2009 he became an associate professor at FAS UWB. The main field of his research and teaching activities is computer speech processing, especially speech synthesis.