MEASUREMENT SCIENCE REVIEW, **17**, (2017), No. 6, 257-263

DE GRUYTER OPEN

MEASUREMENT SCIENCE REVIEW

Journal homepage: http://www.degruyter.com/view/j/msr

ISSN 1335-8871

Two Methods of Automatic Evaluation of Speech Signal Enhancement Recorded in the Open-Air MRI Environment

Jiří Přibil[1], Anna Přibilová[2], Ivan Frollo[1]

[1]*Institute of Measurement Science, SAS, Dúbravská cesta 9, SK-841 04 Bratislava, Slovakia, Jiri.Pribil@savba.sk*
[2]*Institute of Electronics and Photonics, FEE&IT, SUT, Ilkovičova 3, SK-812 19 Bratislava, Slovakia*

The paper focuses on two methods of evaluation of successfulness of speech signal enhancement recorded in the open-air magnetic resonance imager during phonation for the 3D human vocal tract modeling. The first approach enables to obtain a comparison based on statistical analysis by ANOVA and hypothesis tests. The second method is based on classification by Gaussian mixture models (GMM). The performed experiments have confirmed that the proposed ANOVA and GMM classifiers for automatic evaluation of the speech quality are functional and produce fully comparable results with the standard evaluation based on the listening test method.

Keywords: Acoustic noise suppression, magnetic resonance imaging, speech processing, statistical analysis.

1. INTRODUCTION

Construction of 3D articulatory models is necessary for better representation of the human vocal tract function and the subsequent articulatory speech synthesis. For this reason, the audio signal must be recorded simultaneously with the image scanning [1]. The magnetic resonance imaging (MRI) device is used to obtain the vocal tract images of the articulating person that lies in the scanning area while the MR sequence is running [2], [3]. The MRI equipment consists of a gradient coil system producing three orthogonal linear magnetic fields for spatial scanning. The function of these gradient coils is accompanied by an acoustic noise due to rapidly changing Lorentz forces during fast switching inside the weak static field environment [4]. The speech signal recorded under such conditions may be analyzed only if the adequate signal-to-noise ratio is achieved [5]. Several different methods can be used for reduction of the acoustic noise generated in the MRI scanner [6]-[9]. The problem of processing the speech signal in the presence of noise may be solved by various techniques, e.g., the blind source separation by independent component analysis [10]. In our previous research, the noise reduction method was based on the fact that the mentioned acoustic noise of the MRI machine is a periodic signal with its fundamental frequency that may be filtered and processed in the spectral domain [11]-[12].

Objective or subjective criteria can be used for evaluation of enhancement. The subjective ones are based on auditory evaluation by listeners using various categories, such as the mean opinion score, ABX test for comparison of two speech signals with the third one, recognition of expressive speech, annotation of the speech corpus, etc. [13]). The objective approaches for measuring the speech signal quality [14] comprise, for example, evaluation of differences between the speech spectral envelopes [11] or spectral distances [12], etc. These features may be compared and matched using the statistical approaches, like the analysis of variances (ANOVA) [15], [16] or hypothesis tests [17], [18]. The final evaluation in these approaches bears the form of a recognition score that can be obtained by the methods based on artificial neural networks, the nearest neighbor [19], vector quantization classifiers [20], hidden Markov models [21], and support vector machines (SVM) [22]. However, predominantly, the Gaussian mixture models (GMM) [23] are used. The best results are usually achieved by a fusion of different recognition methods, e.g., combination of GMM with SVM used for speaker recognition in the same way as for language recognition [24].

The paper describes the experiments that use the statistical methods based on the ANOVA analysis and the hypothesis tests and, on the other hand, the GMM-based speech quality classifier. Both approaches are used for automatic evaluation of the speech quality after utilization of three different methods of speech enhancement. The motivation of the work was to find an alternative approach to the standard listening tests. It is important in the cases of small audible (or even indiscernible) differences or when their collective realization is problematic, etc. The main advantage of this system is its automatic functioning without human interaction and the possibility of direct numerical matching of the obtained results using the objective comparison criterion.

DOI: 10.1515/msr-2017-0031

2. METHODS

A. Noise Suppression in speech signal

We analyze functionality and successfulness of application of three different methods of the acoustic noise suppression for enhancement of the speech signal recorded during phonation in the MRI environment:

1. The first noise reduction method (further called as *Nsup1*) is based on limitation of the real cepstrum of the noisy speech and clipping the peaks corresponding to the harmonic frequencies of the acoustic noise [11]. This method works well when the basic pitch period of the human voice differs from the repeating period of the running MR scan sequence [12]. In this case, the speech signal with the superimposed noise is recorded by one pick-up microphone.
2. The second tested noise suppression approach (*Nsup2*) uses a subtraction between the short-time spectra of the audio signals recorded by two microphones: the first one recording the speech together with the acoustic noise, the second one recording only the acoustic noise [11].
3. The third method (*Nsup3*) is based on spectral subtraction of the MR scan periodic noise from the same noise superimposed on the speech signal, however, both short-time spectra are estimated from the recording picked-up by the same microphone [12].

The source-filter speech synthesizer with cepstral parameterization of the impulse response of the vocal tract model is used for the reconstruction after the noise suppression in all cases. Each of the applied methods uses different arrangement and practical realization of the recording process as well as the pick-up microphone(s) location [12].

B. ANOVA-based classification of the speech signal

The first part of our speech quality evaluation after application of different methods of noise suppression in the speech signal recorded in the environment of the open-air MRI device working with the weak magnetic field is based on the ANOVA analysis. This approach focuses on testing whether there is a common mean of speech features from several groups. Besides the ANOVA *F*-test giving the ratio of variances between and within groups [16], the hypothesis probability resulting from the Wilcoxon test [25] or the Mann-Whitney U test [26] comparing whether two samples come from identical distributions with equal medians or they do not have equal medians, the Ansari-Bradley hypothesis test [27] is used to specify whether two distributions are the same or they differ in their variances. For a chosen significance level the resulting logical value "0" denotes that the null hypothesis cannot be rejected and the value "1" indicates that it can be rejected.

In the developed classification method the speech spectral properties and prosodic parameters extracted from the clean speech are stored in the database DB_{Orig}, from the speech with MRI noise in DB_{Nfonat}, and from the de-noised speech in the databases $DB_{Nsup1..N}$, treated separately for male and female voices. These speech features and parameters are processed by the one-way ANOVA analysis and then the histograms of the occurrences are calculated – see the block diagram in Fig.1. Three comparison methods are used for each of the speech features and the following parameters are determined:

1. absolute distance between group means of the original speech and the speech enhanced by the methods $D_{OT1\text{-}3}$ after the multiple comparison applied to ANOVA statistical results – see visualization in Fig.2.a),
2. hypothesis probability based on the Ansari-Bradley or the Wilcoxon test,
3. relative RMS distance D_{RMSrel} between the histograms of features extracted from the DB_{Orig} and $DB_{Nsup1\text{-}N}$, as documented by an example in Fig.2.b).

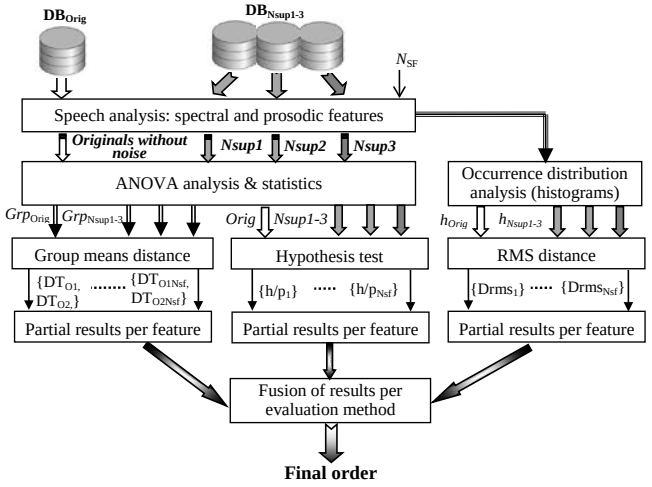

Fig.1. Block diagram of ANOVA-based classifier for evaluation of the enhanced speech signals.

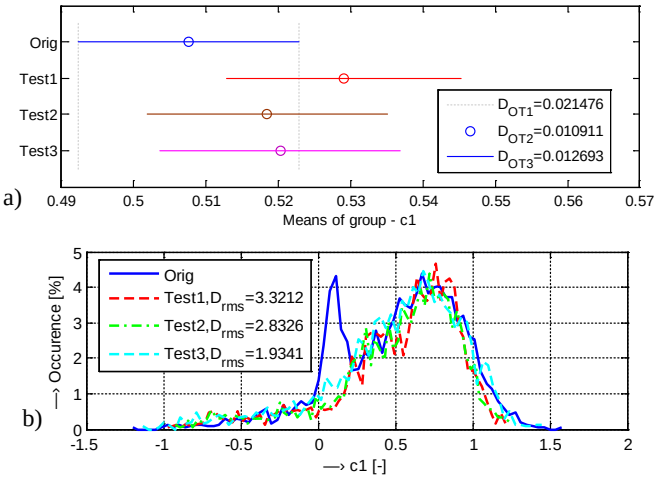

Fig.2. Visualization example of the distance between group means a) and relative RMS distance between the histograms b) for the first cepstral coefficient.

The determined distances and probability values are next sorted by size from minimum (1=nearest to the original) to maximum (3=farthest from the original). From the obtained orders in the range of 1-3 for N_{SF} speech features the histograms of the occurrence distributions are subsequently

calculated for each of the comparison methods (ANOVA/hypothesis test/RMS between histograms) – see the demonstration example in Fig.3. Then, the best order with the maximum occurrence is used for calculation of the final order of mean values for every tested enhanced signal group as the final evaluation value – see the visualization by bar-graphs in Fig.4.

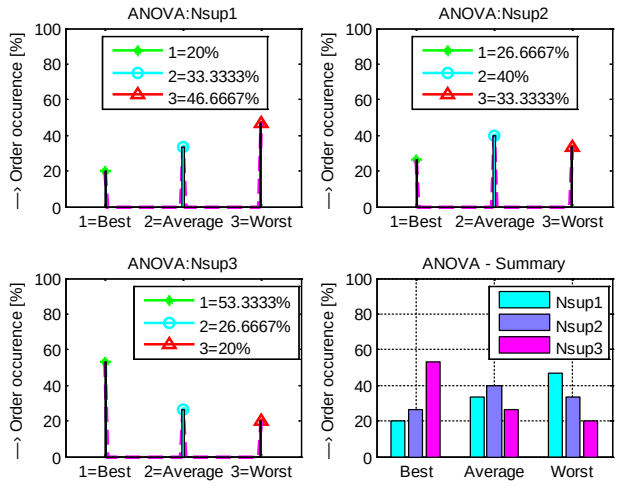

Fig.3. Demonstration example of histograms of order occurences of distances between group means (ANOVA) calculated from all N_{SF} features for three tested noise suppression methods.

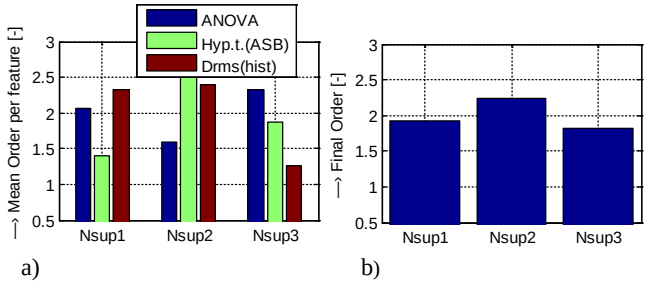

Fig.4. Visualization example of mean values of winner orders per used evalution methods for three tested de-noising approaches a), calculated final order as a result for tested de-noising approaches b).

C. GMM-based evaluation of the speech signal quality

Primarily, the GMMs represent a linear combination of multiple Gaussian probability distribution functions of the input data vector [23]. The covariance matrix and the vector of means together with the weighting parameters have to be determined from the input training data. For the mixture of Gaussians the use of maximum likelihood gives no closed-form analytical solution which would be an ideal case, so the expectation-maximization (EM) iteration algorithm is used for maximizing the likelihood functions [15]. The initial parameters for the EM algorithm are first of all the number of mixtures N_{MIX} and the number of iterations. In general, the elements of the feature vectors could be correlated so that rather a high number of mixture components and a full covariance matrix would be necessary for sufficient approximation. On the other hand, the GMM with a diagonal covariance matrix is usually used in speaker identification [23] due to its lower computational complexity. The GMM classifier returns the probability score that the tested speech signal belongs to the GMM model.

In the standard realization of the GMM classifier, the resulting class is given by the maximum overall probability of all obtained scores (T, n) corresponding to N output classes using the feature vector T from the currently processed speech signal. The main idea of the proposed evaluation method is based on the correlation between the score maxima obtained using the models of the clean speech (further called *Orig*) and the speech with the MRI noise (*Nfonat*). The obtained normalized score values for the enhancement methods *Nsup*1-3 are next ordered using the '*ascend*' sorting for the clean speech models and the '*descend*' sorting for the noisy speech. Finally, the mean score order values in the range of 1-3 (for comparison with the results achieved by the listening tests where "1" represented the best, "2" average, and "3" the worst speech quality) are used for the speech quality evaluation – see an example in Fig.5. The functional block diagram of the whole evaluation method of the speech signal enhancement is shown in Fig.6.

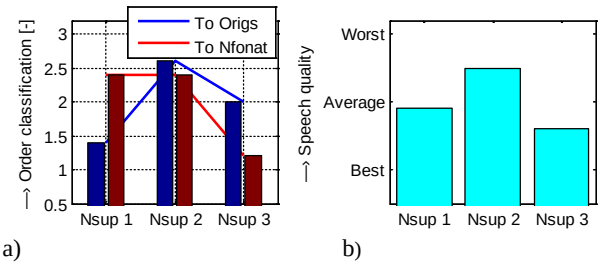

Fig.5. Example of score order determination: partial results for female speaker summarized for all five vowels a), final score as the speech quality evaluation results b).

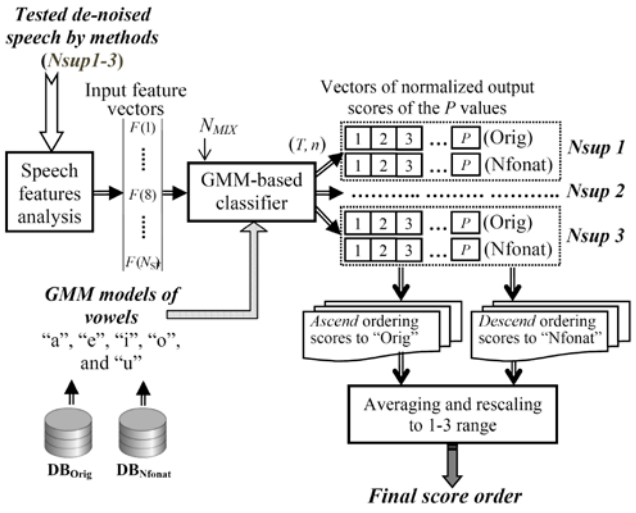

Fig.6. Block diagram of the GMM-based classifier for evaluation of the MRI noise suppression in the speech signal.

D. Determination of features of the speech signal

In the area of the GMM-based speaker [28]-[30], as well as the acoustic signal recognition [31], the most commonly used spectral features are mel-frequency cepstral coefficients together with energy and prosodic parameters. In our experiments the features differ for ANOVA-based evaluation and GMM-based classification of the speech signal quality. The analysis of the input sentence begins with segmentation and determination of the fundamental frequency F0 from the segmented input signal. Next, the smoothed spectral envelope and the power spectral density from the weighted *P* frames of the speech signal are computed for determination of basic and supplementary spectral features. The basic spectral properties are expressed by the statistical parameters as centroid (SC), flatness (SF), spread, skewness, kurtosis, etc. As supplementary spectral features the following parameters are used: spectral decrease (tilt), harmonics-to-noise ratio (HNR), Shannon, Rényi, or Tsallis spectral entropy (SHE/RSE/TSE), etc. For voiced speech description, the first two formant positions F_1, F_2 and their ratio F_1/F_2 are also used in our experiments. The cepstral coefficients $\{c_n\}$ obtained during the process of cepstral analysis, giving information about spectral properties of the human vocal tract, are also successfully used in the feature vectors. The supra-segmental properties include also the speech signal energy expressed by the first cepstral coefficient (En_{c0}) or by the autocorrelation function (En_{r0}). The prosodic parameters consist of two types of energy parameters calculated from the differential microintonation signal $F0_{DIFF}$, zero crossing frequency F_{ZCR}, jitter, and shimmer.

These speech features are stored to different databases depending on the input signal used (DB_{Orig}, DB_{Nfonat}, and $DB_{Nsup1..N}$) – see the block diagram in Fig.7. For the GMM-based experiments, every vector of *P* speech features is subsequently processed to obtain N_{SF} representative statistical values (mean, median, rel. maximum, rel. minimum, etc.).

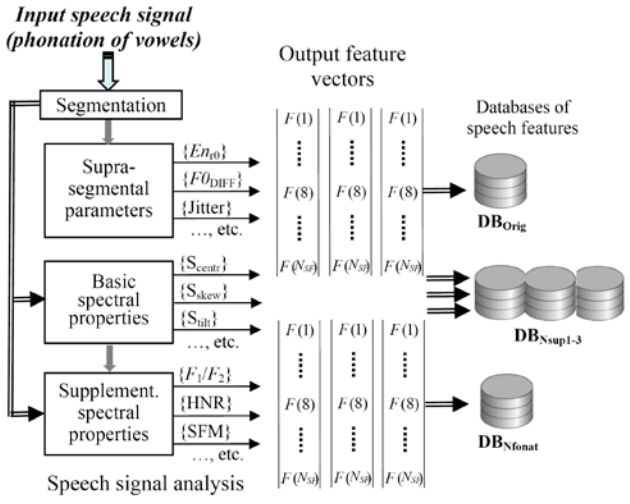

Fig.7. Block diagram of the feature database creation from the speech spectral properties and supra-segmental parameters.

3. MATERIAL AND EXPERIMENTS

A. Speech signal recording and processing

Our experiments were carried out with the open-air MRI equipment E-scan OPERA working with the low magnetic induction of 0.178 Tesla [32]. The speech and noise signals were recorded using the Behringer condenser microphone connected to a separate personal computer via the XENYX 502 mixer and UCA202 audio interface. The audio signals were originally sampled at 32 kHz and then resampled to 16 kHz. The microphone picking up the speech was placed at the position of 150 degrees as documented by the photo of the experimental arrangement in Fig.8. where the tested person lies at 180 degrees. The microphone recording the noise only was placed at 30 degrees. The distance of the microphones from the MRI device central point was 60 cm, and they were situated vertically in the middle between both gradient coils.

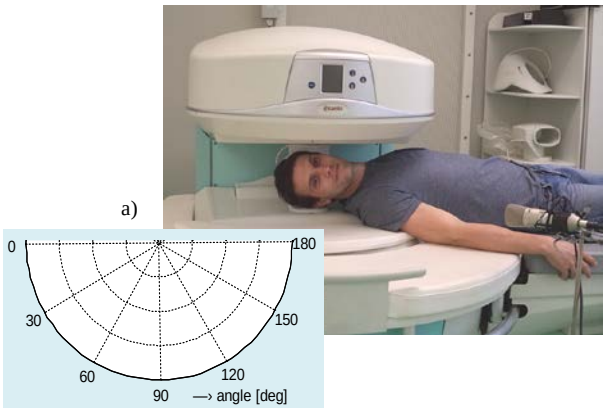

Fig.8. Arrangement of speech and noise recording in the E-scan OPERA: examined person with a pick-up microphone a), principal angle diagram of the MRI scanning area b).

The recorded speech and noise signals were used for creation of the database consisting of five separately phonated long vowels “a”, ”e”, “i”, “o”, and “u” from three male and two female non-professional speakers with time duration interval from 8 to 15 sec. For each of the tested vowels, two types of recordings were carried out. The first one corresponds to the “clean” speech signal of phonation without any MRI noise, only with the superimposed background noise of the temperature stabilizer [12]. The second one is composed of phonation during execution of the MR sequence SSF-3D which is usually applied for MR scanning of the human vocal tract [33].

The input feature vector with the length experimentally set to N_{SF}=16 consisted of a mix of the basic and supplementary spectral and prosodic features. For the ANOVA-based evaluation experiment, the following speech features were used: {En_{c0}, En_{r0}, tilt, SC, flatness, HNR, SHE, RSE, TSE, c_1 – c_3, $F0_{DIFF}$, $F0_{ZCR}$, jitter, and shimmer}. In the case of GMM training and classification the input vector contained statistical representative values of the supra-segmental parameters {$F0_{DIFF}$, jitter, and shimmer}, the basic spectral features determined from the spectral envelopes {F_1/F_2, SC,

tilt}, and the supplementary spectral parameters {HNR, flatness, SHE}.

The Ansari-Bradley hypothesis test was finally used in the ANOVA-based evaluation experiment due to higher consistency of the produced probability results with the absolute distances between group means. In the GMM-based evaluation experiment, a simple diagonal covariance matrix of the GMM as well as the number of mixtures N_{MIX}=8 were finally applied because of their lower computational complexity and relatively good final discriminability of the summary results for all three evaluated methods.

The described analysis and processing of the speech and noise signals were currently realized in the Matlab environment (ver. 2012a), using especially the "Signal Processing" and "Statistics" toolboxes. The Ian T. Nabney "Netlab" pattern analysis toolbox [34] was used for implementation of basic functions for the proposed GMM classifier.

B. Performed evaluation experiments

The subjective evaluation was carried out by the listening test called "*Evaluation of better sound after MRI noise suppression*" by means of the automated internet application located at http://www.lef.um.savba.sk/scripts/itstposl2.dll [35]. This listening test had been accessed by twenty nine listeners in the time period from February 1 to 28, 2017. Our listening test experiment consisted of 10 evaluation sets, each comprising 5 long vowel utterances by male and female voices selected randomly from the speech corpus, so 30 recordings were evaluated in total. For each of the vowel recordings the listener had to choose from four possibilities: "*sounds best*", "*sounds average*", "*sounds worst*" or "*cannot be determined*" – see an example of a screenshot of the listening test in Fig.9. The results obtained in this way are presented in Fig.10.

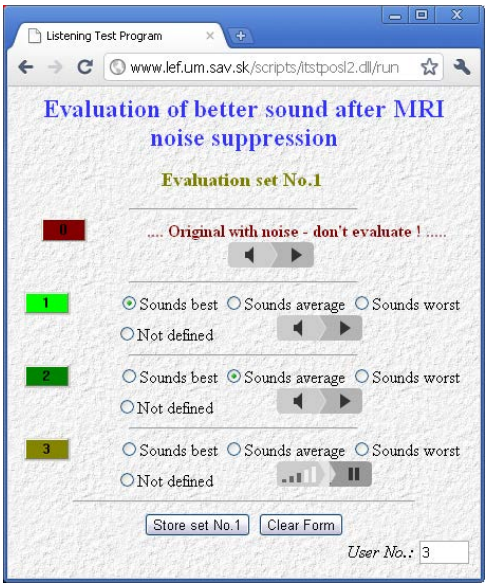

Fig.9. Example of a screen shot of the internet server realization of the listening test; first evaluation set, the first two samples already evaluated, the third one currently playing.

The two basic experiments were focused on verifying the functionality of the developed ANOVA and GMM speech quality classifiers. This step was accompanied by the detailed analysis of the noise suppression method and the speaker type (male/female) – see the partial results in Fig.11. and Fig.12. Finally, the overall obtained values (for all processed vowels and speakers) are matched with the results achieved by the standard listening test method – see Table 2.

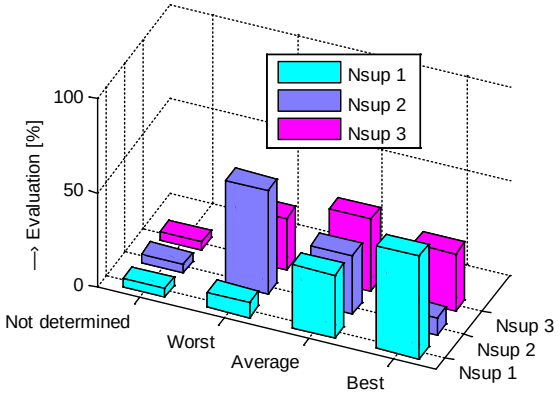

Fig.10. 3D visualization of the evaluation results obtained by the listening test method.

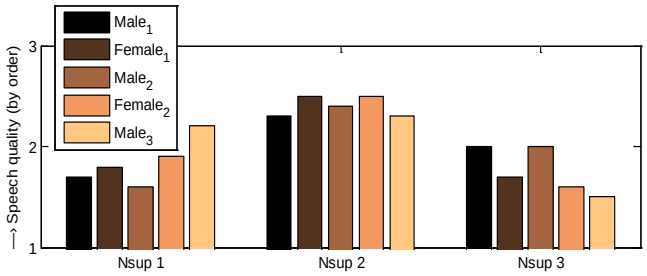

Fig.11. Bar-graph comparison of the final order obtained by the ANOVA evaluation aproach for each of the five tested speakers.

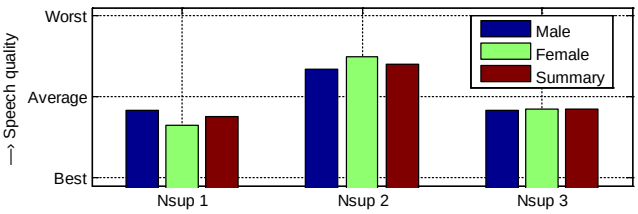

Fig.12. The bar-graph comparison of the GMM-based automatic evaluation results separately for male, female, and both genders of tested speakers, summarized for all five vowels.

Table 2. Final numerical comparison of obtained evaluation orders rescaled to the range of 1-3 (1="*best*", 2="*average*", 3="*worst*").

Method	**Nsup1**	**Nsup2**	**Nsup3**
Listening test	1.52	2.48	1.97
ANOVA-based	1.87	2.34	2.09
GMM-based	1.76	2.41	1.84

4. Conclusion

The performed experiments have confirmed that both proposed automatic classifiers based on statistical approach work correctly and produce results comparable with those attained by the standard listening tests. It was verified on the speech material after the MRI noise suppression consisting of the records of the five basic vowels from five voluntary persons examined in the open-air MRI device during the 3D scanning of human vocal tract.

As documented by the obtained results, the applied setting of the basic parameters for ANOVA and GMM evaluation approaches produces variability of the results for the male/female speakers (see the obtained scores in Fig.11.). On the other hand, the analysis of dependence of the obtained results on different types and different numbers of speech parameters used in the input feature vectors must also be performed. Finally, the computation complexity analysis of the current realization in the Matlab environment revealed that optimization and implementation in a higher programing language is necessary for real-time processing and classification.

Acknowledgment

This work was supported by the Slovak Scientific Grant Agency project VEGA 2/0001/17, the Ministry of Education, Science, Research, and Sports of the Slovak Republic VEGA 1/0905/17, and within the project of the Slovak Research and Development Agency Nr. APVV-15-0029.

The authors would also like to express thanks to all the people who participated in the listening test.

References

[1] Wei, J., Liu, J., Fang, Q., Lu, W., Dang, J., Honda, K. (2016). A novel method for constructing 3D geometric articulatory models. *Journal of Signal Processing Systems*, 82, 295-302.

[2] Aalto, D., Aaltonen, O., Happonen, R.-P. et al. (2014). Large scale data acquisition of simultaneous MRI and speech. *Applied Acoustics*, 83, 64-75.

[3] Kuorti, J., Malinen, J., Ojalammi, A. (2018). Post-processing speech recordings during MRI. *Biomedical Signal Processing and Control*, 39, 11-22.

[4] Tomasi, D., Ernst, T. (2006). A simple theory for vibration of MRI gradient coils. *Brazilian Journal of Physics*, 36, 34-39.

[5] Burdumy, M., Traser, L., Richter, B. et al. (2015). Acceleration of MRI of the vocal tract provides additional insight into articulator modifications. *Journal of Magnetic Resonance Imaging*, 42, 925-935.

[6] Lee, N., Park, Y., Lee, G.W. (2017). Frequency-domain active noise control for magnetic resonance imaging acoustic noise. *Applied Acoustics*, 118, 30-38.

[7] Wu, Z., Kim, Y.C., Khoo, M.C.K., Nayak, K.S. (2014). Evaluation of an independent linear model for acoustic noise on a conventional MRI scanner and implications for acoustic noise reduction. *Magnetic Resonance in Medicine*, 71, 1613-1620.

[8] Oveisi, A., Nestorović, T. (2016). Mu-synthesis based active robust vibration control of an MRI inlet. *Facta Universitatis, Series: Mechanical Engineering*, 14 (1), 37-53.

[9] Sun, G., Li, M., Rudd, B.W. et al. (2015). Adaptive speech enhancement using directional microphone in a 4-T scanner. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 28, 473-484.

[10] Patil, D., Das, N., Routray, A. (2011). Implementation of Fast-ICA: A performance based comparison between floating point and fixed point DSP platform. *Measurement Science Review*, 11 (4), 118-124.

[11] Přibil, J., Horáček, J., Horák, P. (2011). Two methods of mechanical noise reduction of recorded speech during phonation in an MRI device. *Measurement Science Review*, 11 (3), 92-98.

[12] Přibil, J., Přibilová, A., Frollo, I. (2016). Analysis of acoustic noise and its suppression in speech recorded during scanning in the open-air MRI. In *Advances in Noise Analysis, Mitigation and Contro*l. Rijeka, Croatia: InTech, 205-228.

[13] Grůber, M., Matoušek, J. (2010). Listening-test-based annotation of communicative functions for expressive speech synthesis. In *Text, Speech, and Dialogue (TSD) 2010*, LNCS 6231, Springer, 283-290.

[14] Sen, D., Lu, W. (2017). Systems and methods for measuring speech signal quality. *US Patent 9679555*.

[15] Rencher, A.C., Schaalje, G.B. (2008). *Linear Models in Statistics, Second Edition*. John Wiley & Sons.

[16] Lee, C.Y., Lee, Z.J. (2012). A novel algorithm applied to classify unbalanced data. *Applied Soft Computing*, 12, 2481-2485.

[17] Mizushima, T. (2000). Multisample tests for scale based on kernel density estimation. *Statistics & Probability Letters*, 49, 81-91.

[18] Altman, D.G., Machin, D., Bryant, T.N., Gardner, M.J. (2000). *Statistics with Confidence: Confidence Intervals and Statistical Guidelines, 2nd edition*. London: BMJ Books.

[19] Glowacz, A., Glowacz, Z. (2017). Diagnosis of stator faults of the single-phase induction motor using acoustic signals. *Applied Acousticss*, 117, 20-27.

[20] Bapat, O.A., Fastow, R.M., Olson, J. (2013). Acoustic coprocessor for HMM based embedded speech recognition systems. *IEEE Transactions on Consumer Electronics*, 59 (3), 629-633.

[21] Bhardwaj, S., Srivastava, S., Hanmandlu, M., Gupta, J.R.P. (2013). GFM-based methods for speaker identification. *IEEE Transaction on Cybernetics*, 43 (3), 1047-1058.

[22] Vít, J., Matoušek, J. (2013). Concatenation artifact detection trained from listeners evaluations. In *Text, Speech and Dialogue 2013*, LNAI 8082, Springer, 169-176.

[23] Reynolds, D.A., Rose, R.C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3, 72-83.

[24] Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E., Torres-Carrasquillo, P.A. (2006). Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20 (2-3), 210-229.

[25] Rodellar-Biarge, V., Palacios-Alonso, D., Nieto-Lluis, V., Gómez-Vilda, P. (2015). Towards the search of detection in speech-relevant features for stress. *Expert Systems*, 32 (6), 710-718.

[26] Mekyska, J., Janousova, E., Gomez-Vilda, P. et al. (2015). Robust and complex approach of pathological speech signal analysis. *Neurocomputing*, 167, 94-111.

[27] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[28] Venturini, A., Zao, L., Coelho, R. (2014). On speech features fusion, α-integration Gaussian modeling and multi-style training for noise robust speaker classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22 (12), 1951-1964.

[29] Chakroun, R., Zouari, L.B., Frikha, M. (2016). An improved approach for text-independent speaker recognition. *International Journal of Advanced Computer Science and Applications*, 7 (8), 343-348.

[30] Sharma, R., Prasanna, S.R.M., Bhukya, R.K., Das, R.K. (2017). Analysis of the intrinsic mode functions for speaker information. *Speech Communication*, 91, 1-16.

[31] Glowacz, A. (2015) Recognition of acoustic signals of synchronous motors with the use of MoFS and selected classifiers. *Measurement Science Review*, 15 (4), 167-175.

[32] Esaote S.p.A. (2008). *E-scan Opera. Image Quality and Sequences Manual*. 830023522 Rev. A.

[33] Přibil, J., Gogola, D., Dermek, T., Frollo, I. (2012). Design, realization and experiments with a new RF head probe coil for human vocal tract imaging in an NMR device. *Measurement Science Review*, 12 (3), 98-103.

[34] Nabney, I.T. (2004). *Netlab Pattern Analysis Toolbox, Release 3.3*. http://www.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/downloads.

[35] Přibil, J., Přibilová, A. (2013). Internet application for collective realization of speech evaluation by listening tests. In *Proceedings of the International Conference on Applied Electronics (AE2013)*, Plzeň, Czech Republic, 225-228.

Received July 27, 2017.
Accepted November 12, 2017.