# Automatic Text-Independent Artifact Detection, Localization, and Classification in Synthetic Speech

*Jiri PRIBIL [1], Anna PRIBILOVA [2], Jindrich MATOUSEK [3]*

[1] Inst. of Measurement Science, SAS, Dúbravská cesta 9, SK-841 04 Bratislava, Slovak Republic
[2] Slovak University of Technology in Bratislava, IEP FEEIT, Ilkovičova 3, SK-812 19 Bratislava, Slovak Republic
[3] Dept. of Cybernetics, Faculty of Applied Sciences, UWB, Univerzitní 8, CZ-306 14 Plzeň, Czech Republic

jiri.pribil@savba.sk, anna.pribilova@stuba.sk, jmatouse@kky.zcu.cz

Submitted May 24, 2017 / Accepted October 20, 2017

**Abstract.** *The paper describes experiments with statistical approaches to automatic detection, localization, and classification of the basic types of artifacts in the synthetic speech produced by the Czech text-to-speech system using the unit selection method. The first experiment is aimed at artifact detection by the analysis of variances (ANOVA) and hypothesis testing. The second experiment is focused on localization of the detected artifacts by the Gaussian mixture models (GMM). Finally, the developed open-set artifact classifier is described. The influence of the feature vector length and structure on the resulting artifact detection accuracy is analyzed together with other factors affecting the stability of the artifact detection process. Further investigations have shown a relatively great influence of the number of mixtures and the type of a covariance matrix on the artifact classification error rate as well as on the computational complexity. The obtained experimental results confirm the functionality of the artifact detector based on the ANOVA and hypothesis tests, and the GMM-based artifact localizer and classifier. The described statistical approaches represent the alternatives to the standard listening tests and the manual labeling of the artifacts.*

## Keywords

Quality of synthetic speech, analysis of variances (ANOVA), Gaussian mixture models (GMM) classification, text-to-speech (TTS) system

## 1. Introduction

The synthetic speech produced by text-to-speech (TTS) systems is increasingly used to make dialogue management in human-machine interaction more effective. People involved in such a dialogue usually demand high quality, naturalness, and intelligibility of the generated synthetic speech. Various speech synthesis techniques may be implemented in the TTS systems. The most widely used one is the corpus-based speech synthesis using the unit selection (USEL) [1], i.e. selection of the largest suitable segments from the natural speech according to various phonetic, prosodic, and positional criteria, commonly known as the target cost. These speech segments should be smoothly concatenated by minimizing the concatenation cost [2], [3]. However, any concatenation point may become a source of an audible artifact in the finally generated speech [4]. Apart from the wrong description of the natural original speech (such as wrong annotation and/or segmentation [5]), the most dominant causes of the artifacts are related mainly to discontinuities of the fundamental frequency in the voiced speech [6]. From among the other reasons of serious artifacts, time inconsistencies or spectral mismatches at concatenation points can be mentioned [7]. In the process of the TTS system development, all these artifacts must be identified by evaluation methods working without any human interaction. In such an objective method, the automatic speech recognition system yields the final evaluation in the form of a recognition score. Here, the Gaussian mixture models (GMM) [8] are mostly used. In general, the automatic artifact detection, localization, and classification can help in the whole process of the TTS system creation. It holds especially for the artifacts caused by wrong annotation or those found in the already generated synthetic sentence. If their location is known, they can be eliminated in the post-processing or directly during the unit selection as a component part of the concatenation cost.

This work was motivated mainly by the aim of finding an alternative objective approach to the standard listening tests for detection and localization of the artifacts in the synthetic speech. It is important in the cases when the listening test is rather time consuming and relatively difficult due to small audible differences. In addition, the main disadvantage of the human evaluation lies in its subjectivity, lack of reproducibility (different obtained results for repeated tests even from the same subjects), and dependence on ambient conditions. On the other hand, the main advantage of the automatic evaluation system is its function without human interaction and possibility of direct numerical matching of the obtained results using the objective comparison criterion.

The paper describes three basic experiments with developed automatic speech artifact detector, localizer, and

classifier. The functionality of this system is verified and its optimal settings are found. The evaluated objective results are compared with those obtained by the listening test as a subjective rating method.

## 2. Method

Our previous experience with the TTS system based on the USEL synthesis method has shown appearance of six basic types of speech artifacts [9], [10]:

1. local increase of the signal RMS (energy) - $Artf_1$,
2. local decrease of the signal RMS - $Artf_2$,
3. local increase of the fundamental frequency $F_0$ - $Artf_3$,
4. local decrease of $F_0$ - $Artf_4$,
5. superposition of the local energy and the $F_0$ increase/ decrease - $Artf_5$,
6. incorrectly chosen or exchanged speech units - $Artf_6$.

Principally, the proposed artifact automatic detection, localization, and classification system consists of three parts:

- artifact detection based on the analysis of variances (ANOVA) described in the previous paper [9],
- artifact localization developed in concordance with its first stage dealt with in [10] where the position of the GMM score maximum coincides with the location of the artifact inside the tested sentence,
- artifact type classification also based on the GMM approach.

The function of the automatic system begins with the analysis of the tested input sentence. Then, the speech spectral and prosodic features are determined and subsequently applied in the ANOVA detector block making a decision whether a speech artifact is present or not. In this step, the database of the clean synthetic speech ($DB_{CLEAN}$) and the database of the synthetic speech with artifacts ($DB_{ARTF}$) are used – see the block diagram in Fig. 1.

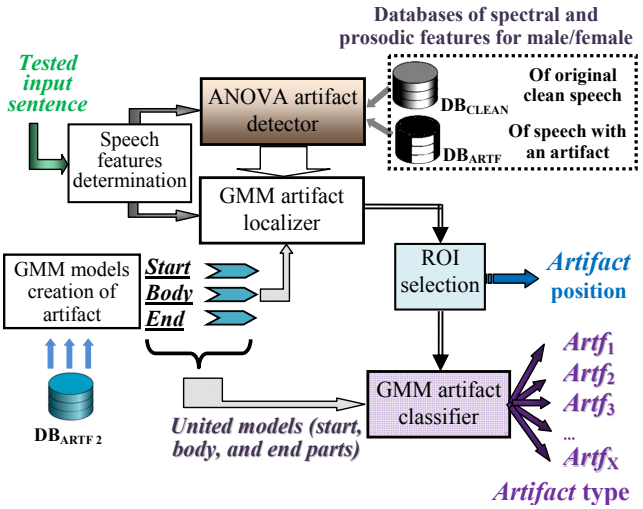


**Fig. 1.** Basic block diagram of the proposed artifact detection, localization, and classification system.

If the sentence is marked as having an artifact, other types of spectral and prosodic parameters are used for artifact localization using the trained GMM models of the starting/ ending parts and the bodies of the artifacts (database $DB_{ARTF2}$). Once the artifact is localized, the nearest region of interest (ROI) is determined and the united GMM models of the starting, ending, and body parts are used for the final classification of the artifact type (see the example in Fig. 2).

### 2.1 Determination of Speech Spectral Features and Prosodic Parameters

The speech artifact detection method begins with listening of the speech signal and its evaluation using the standard audio software or the program system dedicated to speech processing, e.g. Praat [11]. After detection of an audible artifact by repeated listening, the next step is visual evaluation of the speech signal. In this way, the original speech material is selected and prepared for building of the basic speech feature databases $DB_{CLEAN}$ and $DB_{ARTF}$.

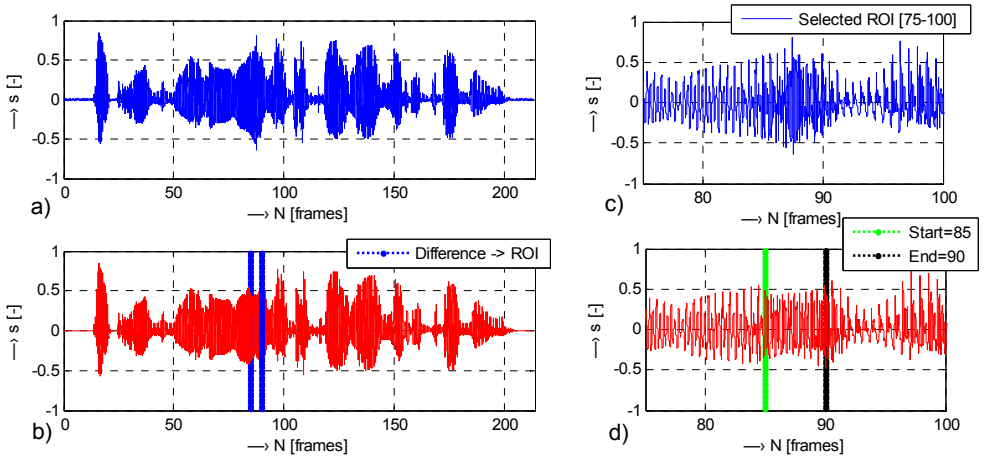


**Fig. 2.** Demonstration of differences in speech signals: the clean sentence "send04g" (a), the sentence with an artifact "send04b"and its ROI (b), detail of the clean signal in the ROI (c), detailed part in the artifact neighborhood with the determined start/end locations (d).

Spectral features like the mel-frequency cepstral coefficients together with the energy and the prosodic parameters are mostly used in the GMM-based speaker identification or verification [12], [13]. Among the other spectral properties, e.g. the first five formants can be used in psychological stress detection in speech [14]. In our experiments the features differ for the ANOVA detection and the GMM localization/classification of the artifacts. In general, three types of speech features can be determined:

1. supra-segmental parameters – speech signal energy calculated from the first cepstral coefficient $c_0$ ($\mathrm{En_{c0}}$) or by the autocorrelation coefficient $r_0$ ($\mathrm{En_{r0}}$), differential $F0$ microintonation ($F0_{\mathrm{DIFF}}$), jitter, shimmer, zero-crossing period ($L_{\mathrm{ZCR}}$), and frequency ($F_{\mathrm{ZCR}}$).
2. basic spectral features – first two formants ($F_1$, $F_2$), their ratio ($\mathrm{F_1/F_2}$), spectral tilt ($S_{\mathrm{tilt}}$), spectral centroid ($S_{\mathrm{centr}}$), statistical measures describing the spectral shape: spectral spread ($S_{\mathrm{spread}}$), skewness ($S_{\mathrm{skew}}$), and kurtosis ($S_{\mathrm{kurt}}$).
3. supplementary spectral features – harmonics-to-noise ratio (HNR), spectral flatness measure (SFM), Shannon spectral entropy (SHE), Rényi spectral entropy (RE), and Tsallis spectral entropy (TE).

The determined speech features are structured as vectors with the length $N_{\mathrm{SF}}$ and stored in the $\mathrm{DB_{CLEAN}}$ and $\mathrm{DB_{ARTF}}$ databases separately for male and female voices. The $\mathrm{DB_{ARTF2}}$ database comprises the separate speech features determined from the start, end, and body parts of the artifacts for localization and classification. For the GMMs creation and training, the representative statistical values (mean, median, rel. maximum, rel. minimum, skewness, kurtosis, etc.) are calculated from the original speech features.

To obtain the relevant speech features from $\mathrm{DB_{CLEAN}}$ and $\mathrm{DB_{ARTF}}$ databases the criterion of mutual independence between the synthetic speech with and without artifacts is applied. The final value of the mutual independence for every feature and every category is evaluated using three parameters:

1. relative RMS distance $D_{\mathrm{RMSrel}}$ between the histograms of features extracted from the $\mathrm{DB_{CLEAN}}$ and $\mathrm{DB_{ARTF}}$,
2. absolute distance between group means $D_{12}$ after the multiple comparison of the group means applied to ANOVA statistical results,
3. hypothesis probability resulting from the Wilcoxon test [15] or the Mann-Whitney U test [16] comparing whether two samples come from identical distributions with equal medians or they do not have equal medians.

For all the three parameters, the features are sorted, in such a way that the higher the index, the lower the mutual independence. The parameter quantifying the mutual independence of the databases for every feature ($MUTI_{\mathrm{SF}}$) is represented by the resulting mean position in the category calculated as

$$MUTI_{\mathrm{SF}} = \begin{cases} \mathrm{Mean}\ \left(cw_1 \cdot cr_1,\ cw_2 \cdot cr_2,\ cw_3 \cdot cr_3\right) & h = 1 \\ N_{\mathrm{SFC}} & h = 0 \end{cases} \quad (1)$$

where $N_{\mathrm{SFC}}$ is the number of features in the category, $cr_1$ is a criterion by $D_{\mathrm{RMSrel}}$, $cr_2$ by $D_{12}$, $cr_3$ by the hypothesis probability, and $cw_{1\text{-}3}$ are individual weights depending on the importance of the criterion. If the null hypothesis cannot be rejected for any feature, it is penalized by the highest index of the sorted vector $N_{\mathrm{SFC}}$. The features are selected by two rules: exclusion of the features with very small null hypothesis probabilities, and elimination of those with small RMS distances between the speech with and without artifacts. This feature separation process is performed with the speech material comprising sentences spoken by all speakers.

## 2.2 Artifact Detection Based on ANOVA

The first step of our speech artifact identification experiment based on ANOVA analysis is focused 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 we use the Ansari-Bradley probability test specifying whether two distributions are the same or they differ in their variances. For

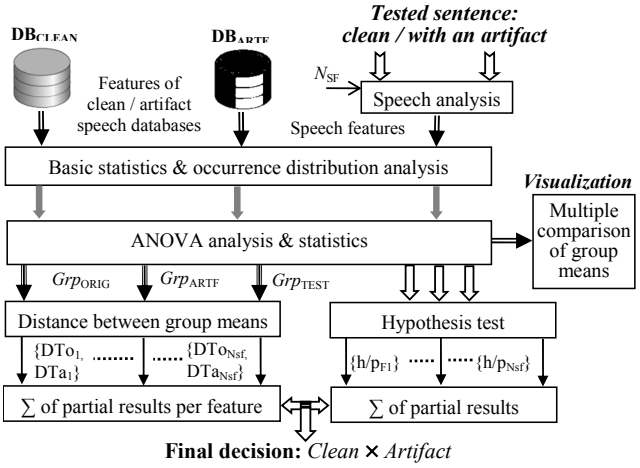


**Fig. 3.** Block diagram of the developed classifier using comparison of group means and the hypothesis test from the ANOVA statistics.

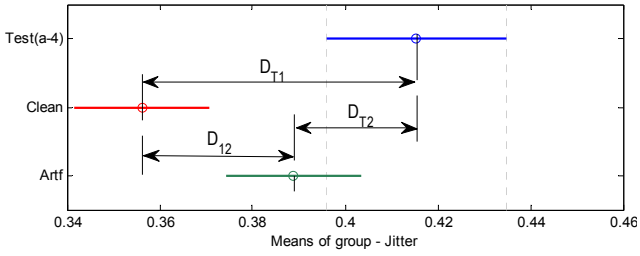


**Fig. 4.** Multiple comparison of group means of ANOVA results for jitter. Distances: $D_{\mathrm{T1}}$ – tested and clean sentences (0.06), $D_{\mathrm{T2}}$ – tested sentences and those with artifacts (0.02), $D_{12}$ – sentences with and without artifacts (0.04); male sentence "send04b" with an artifact.

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. The overall structure of the method can be seen in Fig. 3.

The speech spectral properties and prosodic parameters obtained during analysis of the tested sentence are used to calculate the corresponding basic statistical parameters and the occurrence distributions of the feature values. Further, they are processed by the one-way ANOVA analysis. The distances between the means of the groups are visualized using the multiple comparisons of groups (see Fig. 4). The minimum absolute value of the group distance is found from among the distances between the group means:

- $D_{T1}$ – the tested sentence and the clean sentence,
- $D_{T2}$ – the tested sentence and the one with an artifact,
- $D_{12}$ – the clean sentence and the one with an artifact.

For each of $N_{SF}$ speech features these results yield the decision about the tested sentence (clean/artifact), and the Ansari-Bradley test between probability distributions gives the probability and the logical output value (0/1).

## 2.3 GMM-based Artifact Localization

The GMMs represent a linear combination of multiple Gaussian probability distribution functions of the input data vector. For their creation it is necessary to determine the covariance matrix, the vector of means, and the weights from the input training data. In general, spherical, diagonal, or full covariance matrices may be used. If the elements of the feature vectors are correlated, their number must be relatively high and satisfactory approximation can be achieved only with the full covariance matrix. However, in the speaker identification tasks, the diagonal covariance matrix is used due to its lower computational complexity. The maximum likelihood function of the GMM is found by the expectation-maximization iteration algorithm. It is controlled by the number of mixtures $N_{MIX}$ and the number of iterations. The classifier returns the probability that the tested utterance belongs to the GMM model. In the standard GMM classifier, the resulting class is given by the maximum overall probability of all the obtained scores corresponding to $K$ output classes. Here, only one output class is defined and the GMM classifier processes $N$ feature vectors corresponding to $N$ frames of the tested sentence.

The main idea of the proposed localization method is based on the assumption of correlation between the position of the artifact and the score maximum from the vector of normalized scores obtained by comparison between the currently tested speech frame and the trained GMM model (see Fig. 5). Three types of the GMM models of the artifacts are created and trained for each voice:

a) starting part – speech signal in the left margin frame of the artifact and ±$i$ frames in its neighborhood,

b) ending part – speech signal in the right margin frame of the artifact and ±$i$ frames in its neighborhood,

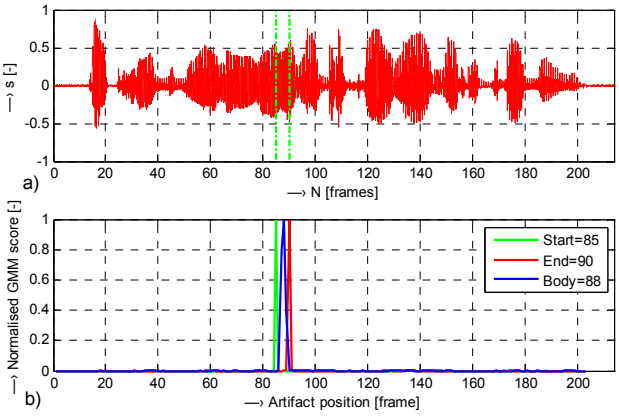


**Fig. 5.** Demonstration of artifact localization using 3 GMM models for start, end, and body parts: sentence "sent04b" with a manually determined artifact (a), normalized GMM scores for 3 models (b).

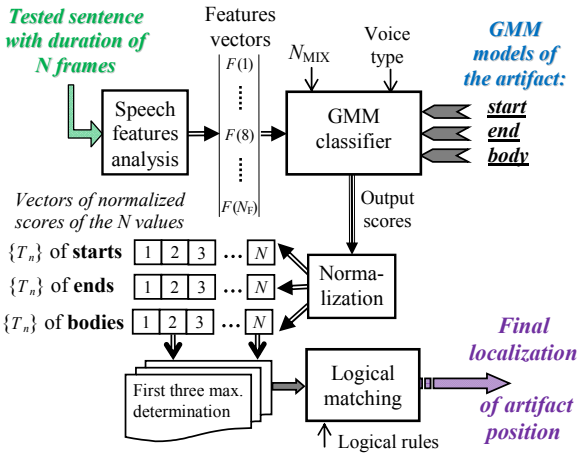


**Fig. 6.** Block diagram of the GMM classifier for artifact localization.

c) body of the artifact – speech signal spanning from the starting to the ending frame.

In the classification phase, the input feature vectors are compared with these 3 trained GMM models to get 3 output vectors of normalized scores. For final localization of the artifact position the first 3 maxima are evaluated by logical matching with the predefined rules (see Fig. 6) covering the situations when the localization algorithm might fail – the starting frame position must precede the ending one, the artifact body must lie between the start and the end, etc. If one of these conditions is not fulfilled, the position will be assigned to the 2nd or the 3rd determined score maximum. Only one artifact within the tested sentence can be found by this approach and the artifact presence must be confirmed by another detection method, e.g. ANOVA-based approach.

## 2.4 Classification of Artifact Types

The artifact types 1-5 occur relatively often, so the corresponding changes of prosodic and spectral parameters may be defined appropriately and the classification can be carried out with a relatively high precision. In the last class,

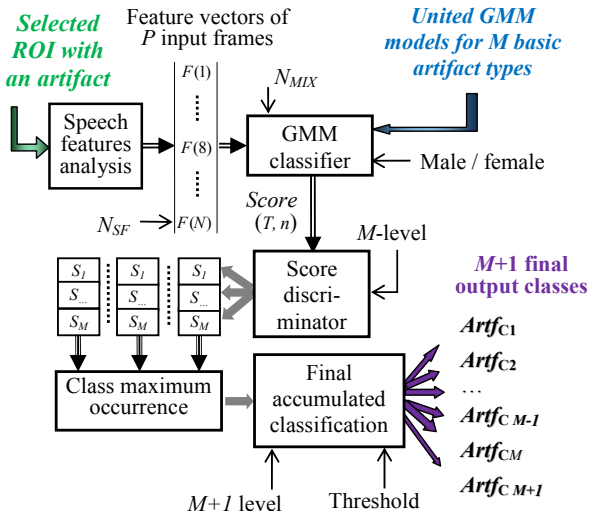


**Fig. 7.** Block diagram of the GMM-based open-set classifier for artifact identification after its detection and localization.

the amount of reference data is not sufficient for the GMM model creation and training due to a different context of an artifact each time it appears. Therefore, the GMM-based classifier must be created in the open set with the 6th class containing all artifacts that had not been classified as the types 1-5.

The last part of our experiment begins with training of the united GMM models on speech signals of the start, body, and end of the artifact with ±*i* frames in the left/right vicinity of ROI. In the classification phase, the input feature vectors from the tested sentence are compared in parallel with 3 trained GMMs to obtain 3 output vectors of the normalized scores. These output scores are analyzed to determine the maximum overall probability in the discriminator block performing basic classification to one of *M* output classes assigned to each of the processed speech feature vectors (see Fig. 7). Next, the class distribution based on histograms is constructed and the maximum occurrence is determined. The final classification block works with *M*+1 output classes – the virtual class is added to the basic closed set of *M* artifact types to create the open-set artifact identifier. The classification strategy is based on the consideration that when the class distribution has no dominant class, the whole tested sentence finally belongs to the 6th class. Practically, the maximum occurrence is compared with the threshold $Tresh_0$ given as a ratio between the number of the currently processed frames and the number of the basic classes (*M*).

# 3. Material, Experiments, and Results

Three basic comparison experiments were performed within the research described in this paper: the first one is the verification of functionality of the ANOVA-based artifact detector using the synthetic speech produced by the Czech TTS system. The second experiment compares the automatically localized artifact position using the GMM-based classifier. The third experiment consists in testing and verifying of the proposed automatic GMM-based artifact type classifier.

The correctness of selection of the ROI with the artifact inside the tested sentence was checked for its influence on the accuracy and the stability of the classification results. In the auxiliary experiments we analyze the influence of different types of speech spectral features and suprasegmental parameters on the resulting artifact detection accuracy. Next, the localization accuracy is analyzed using the artifact position relative error ($APE_{rel}$) and then compared regarding the number of used GMM mixture components. Furthermore, the dependence of the error rate of artifact classification (ERAC) on the number of GMM components and on the method of the covariance matrix calculation was analyzed. Finally, the computational complexity (CPU processing time) was evaluated with the aim to find critical parts of the proposed algorithms and subsequently to make an optimization for real-time processing.

## 3.1 Material and Processing Conditions

The artifact detection, localization, and classification experiments use the synthetic speech produced by the Czech TTS system implementing the USEL synthesis method [17–19]. The main speech corpus was divided into two parallel groups of 40 declarative sentences of the male/female voices. The first group comprises the sentences without any audible artifact designated as "clean"; the second group consists of the same sentences produced by the same male and female TTS voices with just one speech artifact in each sentence. All the sentences with duration 2.5 to 5 s were sampled at 16 kHz. The derived database of ROIs of the artifacts was used for training of the GMM models to classify the artifact types. Independence of the male/female voices during the training and the testing was achieved by the data *k*-fold cross-validation. The groups of sentences were divided by the ratio of 3:1 – three for the training and one for the testing/classification. Due to a limited number of sentences with "real" artifacts occurring during the TTS synthesis, the classical cross-validation data selection could not be used in the GMM-based artifact localization. Therefore, for the testing in the localization experiment, another 20 + 20 sentences with artifacts were derived by cutting or adding a sentence part, using a signal from another sentence, etc. to change the position of the artifact in the sentence.

For determination of the $MUTI_{SF}$ values 25 different types of speech features were tested: 10 prosodic parameters, 10 basic spectral features, and 5 supplementary spectral features (see the detailed results for the prosodic features in Fig. 8a, the basic spectral features in Fig. 8b, the values for the supplementary spectral ones in Tab. 1). Due to statistical similarity between the "*clean*" and "*artifact*" groups, 10 features with the lowest mutual independence were omitted, so 15 speech features were used for ANOVA -based detection. In accordance with the previous research [10] the basic classification of artifacts in the speech utilizes 6 feature sets of 9 items (Tab. 2).

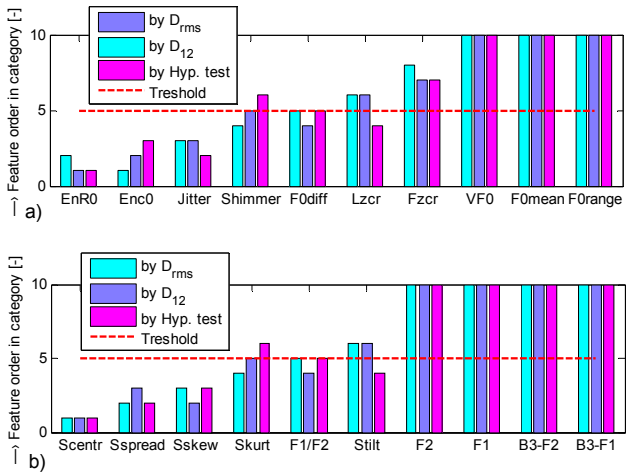


**Fig. 8.** Feature positions in sorted vectors evaluated by $D_{RMSrel}$, $D_{12}$, and the probability of rejection of the null hypothesis for prosodic (a) and basic spectral features (b); for all voices, features with the zero hypothesis test value were further automatically omitted; the threshold was equal to 5.

| Feature | $cr_1$ order | $cr_2$ order | $cr_3$ order | $MUTI_{SF}$ [-]* | Hypothesis** |
|---|---|---|---|---|---|
| HNR | 1 | 2 | 1 | 1.22 | 1 |
| SHE | 2 | 1 | 2 | 1.43 | 1 |
| SFM | 3 | 3 | 3 | 2.65 | 1 |
| RE | 5 | 5 | 5 | $N_{SFC}$ | **0** |
| TE | 5 | 5 | 5 | $N_{SFC}$ | **0** |

* $N_{SFC} = 5$; $cw_1 = 0.75$, $cw_2 = 1$, $cw_3 = 0.9$.
** For 5 % significance level, features with $h = 0$ were finally omitted, the threshold was equal to 3.

**Tab. 1.** Detailed results of the mutual independence and the hypothesis values for the supplementary spectral category for all voices.

| | P0 | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|---|
| $SF_1$ | $En_{r0}$ | $En_{r0}$ | $En_{r0}$ | $En_{r0}$ | $En_{r0}$ | $En_{r0}$ |
| $SF_2$ | $En_{c0}$ | $En_{c0}$ | $En_{c0}$ | $En_{c0}$ | $S_{centr}$ | $S_{centr}$ |
| $SF_3$ | HNR | $S_{spread}$ | $F_1/F_2$ | HNR | SFM | HNR |
| $SF_4$ | $S_{centr}$ | $S_{tilt}$ | HNR | $S_{centr}$ | SHE | SHE |
| $SF_5$ | SFM | SFM | $F_{ZRC}$ | SFM | $F_{ZCR}$ | SFM |
| $SF_6$ | SHE | SHE | $L_{ZCR}$ | SHE | $L_{ZCR}$ | $S_{spread}$ |
| $SF_7$ | $F0_{DIFF}$ | $F0_{DIFF}$ | $F0_{DIFF}$ | $S_{skew}$ | $F0_{DIFF}$ | $S_{skew}$ |
| $SF_8$ | $J_{abs}$ | $J_{abs}$ | $J_{abs}$ | $F_1/F_2$ | $J_{abs}$ | $S_{kurt}$ |
| $SF_9$ | $AP_{rel}$ | $AP_{rel}$ | $AP_{rel}$ | $AP_{rel}$ | $AP_{rel}$ | $S_{tilt}$ |

**Tab. 2.** Internal structure of the tested speech feature sets P0-P5. Only features with $MUTI_{SF} < N_{SFC}$ and $h = 1$ were used.

Influence of different number of features with high mutual independence between the synthetic speech with and without artifacts was analyzed for three feature vector lengths: 5, 9, 15 (PN5, PN9, PN15). The shortest one PN5 consists of the features with the first five smallest $MUTI_{SF}$: $En_{r0}$, HNR, $F0_{DIFF}$, jitter, shimmer. The second one PN9 includes also the features with the $MUTI_{SF}$ value below the threshold containing the features of the set P3 for the male voice and P4 for the female voice. The extended vector PN15 consists of the features with $MUTI_{SF} < N_{SFC}$ and $h = 1$: $En_{c0}$, $En_{r0}$, HNR, $S_{centr}$, $S_{spread}$, $S_{skew}$, $S_{tilt}$, SFM, SHE, $F0_{DIFF}$, $L_{ZCR}$, $F_{ZCR}$, $F_1/F_2$, $J_{abs}$, $AP_{rel}$. According to the results published in [10], [20], the length of the input data vector for GMM training and testing was set to 16.

The objective ANOVA-based evaluation was performed separately for each gender. The resulting artifact detection accuracy was calculated from the number $X_a$ of correctly identified artifact/clean sentences and the total number $N_u$ of sentences as $(X_a/N_u) \times 100$ [%]. The artifact neighborhood before its beginning and after its end was set to ±11 frames in correlation with [10]. The artifact position relative error $APE_{rel}$ in frames was calculated as the average of the absolute position error of the starting and the ending parts $APE_{ABSstart}$, $APE_{ABSend}$ in every sentence as

$$APE_{rel} = \frac{APE_{ABSstart} + APE_{ABSend}}{2} \cdot w_O \quad [\text{frames}] , \quad (2)$$

where $w_O$ is the frame shift for analysis chosen as one fourth of the frame in samples. To determine the dominant class inside the open-set classification the threshold was set experimentally to $1.2 \times Tresh_0$ (i.e. adding 20 % to the basic level given by the calculated $P/M$ ratio). In all cases, the ROI was selected manually for further comparison and evaluation of the $ERAC$ calculated from the number $X_C$ of the sentences with the correctly determined artifact class and the total number $N_T$ of the tested sentences as

$$ERAC = \left(1 - \frac{X_C}{N_T}\right) \cdot 100 \quad [\%] . \quad (3)$$

The described speech signal processing was realized in the Matlab environment (ver. 2012a) and the basic functions of the Nabney "Netlab" pattern analysis toolbox [21] were used in the GMM classifier. The computational complexity was determined using the UltraBook with the following configuration: processor Intel(R) Intel i5-4200U at 2.30 GHz, 8 GB RAM, and Windows 10 (64-bit) OS.

## 3.2 ANOVA-Based Artifact Detection

For verification of functionality of the ANOVA-based artifact detector the subjective artifact determination was performed using the conventional listening test "*Synthetic speech quality evaluation – male / female voice*" by the automated internet application located on the web page http://www.lef.um.savba.sk/scripts/itstposl2.dll. It had been accessible from June 15 to 30, 2014 and then the results were processed. Twenty one listeners (5 women, 16 men) took part in this subjective evaluation consisting of 42 listening tests (21 male, 21 female voices). This internet application in the form of MS ISAPI/NSAPI DLL script runs on the server PC and communicates with the user within the framework of the HTTP protocol by means of HTML pages. The complete test consists of 10 evaluation sets with random selection of sentences. For each sentence there is a choice from three possibilities: "*clean - without artifact*", "*with an artifact*", or "*other - cannot be recognized*". The resulting confusion matrix of the results for the male/female voices is shown in Tab. 3, comparison of the artifact detection accuracy based on ANOVA and the lis-

tening test can be seen in Fig. 9. The results obtained from the performed listening tests show principally high successfulness in the subjective evaluation of the synthetic speech artifacts. Particularly, the best results are achieved in the case of the male voice (approx. 95%) in comparison with the accuracy of 89% for the female voice.

| Male | Clean | Artifact | Other |
|---|---|---|---|
| Clean | 94.28 | 2.86 | 2.86 |
| Artifact | 3.81 | 95.24 | 0.95 |

| Female | Clean | Artifact | Other |
|---|---|---|---|
| Clean | 79.05 | 12.38 | 8.57 |
| Artifact | 1.91 | 97.14 | 0.95 |

**Tab. 3.** Confusion matrices of evaluation of the listening test results of clean/artifact sentences for the male and female voices separately.

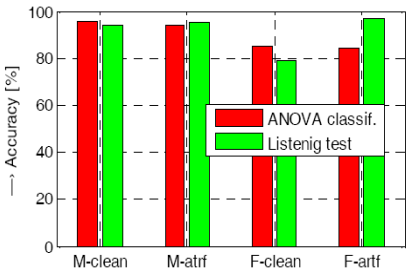


**Fig. 9.** Bar-graph comparison of the mean clean/artifact recognition accuracy based on ANOVA and the listening test (b) for the male (M) and the female (F) voice gender.

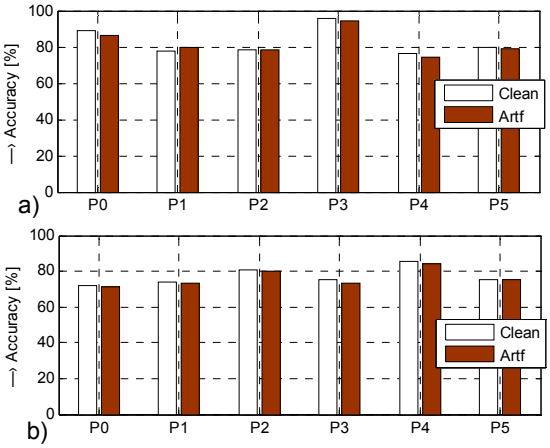


**Fig. 10.** Summary results of the clean/artifact speech detection accuracy for different feature sets: results for male (a), and female (b) voices.

| Feature vector type | Mean artifact detection accuracy [%] | | |
|---|---|---|---|
| | $\text{Clean}_{\text{male/female}}$ | $\text{Artf}_{\text{male/female}}$ | $\text{Sum}_{\text{male/female}}$ |
| PN5 | 76.0 / 70.0 (27.4 / 28.3) | 74.7 / 68.0 (32.8 / 30.3) | **75.3/69.0** |
| PN9 | 95.6 / 85.6 (13.3 / 18.2) | 94.4 / 84.4 (16.7 / 21.4) | **95.0/85.0** |
| PN15 | 80.0 / 72.0 (26.9 / 27.8) | 76.0 / 70.7 (28.5 / 29.1) | **78.0/71.3** |

**Tab. 4.** The mean artifact detection accuracy (its standard deviation in parentheses) for different number of used speech features; results determined for male/female voices.

In this part of the experiment, the following two auxiliary investigations were performed:

1. effect of the feature vector composition (sets P0-P5) on the clean/artifact detection accuracy (Fig. 10),
2. effect of the number of used features on the mean clean/artifact speech detection accuracy (Tab. 4).

Figure 10 shows that, for the male voice, the highest accuracy (94%) was achieved for the set P3 consisting of all three feature categories. The best accuracy (84%) for the female voice corresponds to a different mix of these three feature categories (set P4) being contrariwise almost the worst for the males. Generally, the artifact detection in the sentences was more successful for the male voice than for the female one.

It might be caused by not finding the proper features for classification (in principle, for the female voice there is higher variability on the supra-segmental as well as on the spectral level). The second auxiliary ANOVA-based experiment documents that the accuracy is greatly affected by limitation of the speech feature vector length $N_{SF}$. Higher error rates were produced for the numbers of features lower than 9, however, for 15 features there is not adequate impact on the artifact detection accuracy as documented in Tab. 4.

## 3.3 GMM-Based Artifact Localization

The second basic experiment compares automatically localized artifact positions using the GMM-based classifier. The positions determined manually by the Praat program and by the listening were used to calculate $APE_{\text{rel}}$. In addition, two auxiliary experiments were realized with the aim to cover:

1. effect of the number of mixtures $N_{\text{MIX}} = \{16, 32, 48, 64, 128\}$ during GMM training on $APE_{\text{REL}}$ (Fig. 11),
2. the computational complexity: CPU times of the GMM training and classification phases for different number of mixtures (Fig. 12).

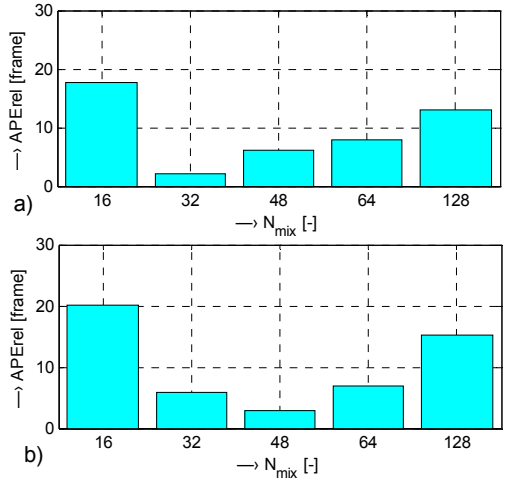


**Fig. 11.** Final comparison of the $APE_{\text{REL}}$ values depending on the used number of GMM mixtures for male (a) and female (b) voices.

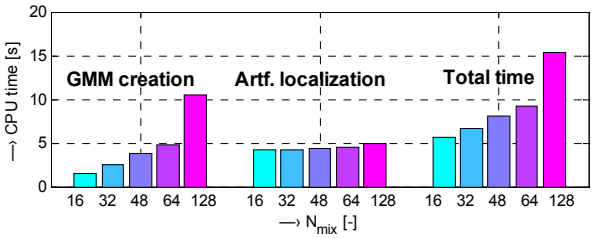


**Fig. 12.** Comparison of the computational complexity (CPU time in [s]) for different number of used GMMs in the artifact localization experiment; $G_{CMtype}$= Full.

The obtained results in this part of the experiment document proper functioning of the developed GMM-based artifact localizer. The analysis has shown principal impact of different number of mixtures on the localization precision (see the bar-graph comparison in Fig. 11). For this reason, a suboptimum of 32/48 mixtures was finally applied. The complexity of computation depends on the number of applied mixtures only in the creation and training of the GMMs, not in the localization/classification. The bar-graph in Fig. 12 shows that the computation time is 10 times higher for 128 mixtures than for 16 mixtures, however, unexpectedly, according to Fig. 11 the accuracy of fixing the artifact position decreases for more than 32/48 mixtures in the case of the male/female voice.

## 3.4 GMM-Based Artifact Classification

The third experiment consists in testing and verifying whether the proposed automatic GMM-based classifier of artifacts is principally correct and produces sufficiently low *ERAC* – see the best results in the form of the confusion matrices in Tab. 5, 6 for the male/female voices. Then the influence of selection of ROI with an artifact in the tested sentence was compared (see the results in Tab. 7). Three auxiliary experiments were realized to analyze:

1. influence of the number of mixtures $N_{MIX}$ = {4, 8, 16, 32} on *ERAC* – see Tab. 8,
2. the method of covariance matrix calculation during GMM training: $G_{CMtype}$={'*Spherical*', '*Diagonal*', '*Full*'} – compare *ERAC* values in Tab. 9,
3. the computational complexity: CPU times for the GMM training and classification using different types of covariance matrices for all voices (Tab. 10).

The obtained results show that, if the ROI is not set and the whole sentence is analyzed, the error rate will be unacceptable especially for the virtual 6th class – compare

| | $Artf_{C1}$ | $Artf_{C2}$ | $Artf_{C3}$ | $Artf_{C4}$ | $Artf_{C5}$ | $Artf_{C6}$ |
|---|---|---|---|---|---|---|
| $Artf_{C1}$ | **100** | 0 | 0 | 0 | 0 | 0 |
| $Artf_{C2}$ | 0 | **100** | 0 | 0 | 0 | 0 |
| $Artf_{C3}$ | 0 | 0 | **100** | 0 | | 0 |
| $Artf_{C4}$ | 0 | 0 | 0 | **100** | 0 | 0 |
| $Artf_{C5}$ | 0 | 0 | 0 | 0 | **100** | 0 |
| $Artf_{C6}$ | 0 | 20 | 0 | 20 | 0 | **60** |

**Tab. 5.** Confusion matrix of the artifact type GMM-based evaluation in [%]; male voice, $N_{MIX}$ = 16, $G_{CMtype}$= Full.

| | $Artf_{C1}$ | $Artf_{C2}$ | $Artf_{C3}$ | $Artf_{C4}$ | $Artf_{C5}$ | $Artf_{C6}$ |
|---|---|---|---|---|---|---|
| $Artf_{C1}$ | **100** | 0 | 0 | 0 | 0 | 0 |
| $Artf_{C2}$ | 0 | **100** | 0 | 0 | 0 | 0 |
| $Artf_{C3}$ | 0 | 0 | **100** | 0 | 0 | 0 |
| $Artf_{C4}$ | 0 | 0 | 0 | **100** | 0 | 0 |
| $Artf_{C5}$ | 0 | 50 | 0 | 0 | **50** | 0 |
| $Artf_{C6}$ | 0 | 0 | 0 | 0 | 0 | **100** |

**Tab. 6.** Confusion matrix of the artifact type GMM-based evaluation in [%]; female voice, $N_{MIX}$ = 16, $G_{CMtype}$ = Full.

| **Selection / *ERAC* [%]** | $Artf_1$ | $Artf_2$ | $Artf_3$ | $Artf_4$ | $Artf_5$ | $Artf_6$ | ***Mean*** |
|---|---|---|---|---|---|---|---|
| ROIs – correct | 0 | 0 | 0 | 0 | 25 | 20 | **7.5** |
| Full sentences | 62 | 51 | 57 | 48 | 72 | 85 | **62.5** |
| ROIs- incorrect | 100 | 50 | 100 | 100 | 80 | 50 | **80** |

**Tab. 7.** Dependence of mean ERAC on correctness of ROI selection in the tested sentence; for all voices, $N_{MIX}$ = 16, $G_{CMtype}$ = Full.

| **Voice / *ERAC* [%]** | **$N_{MIX}$ ($G_{CMtype}$ = Full)** | | | |
|---|---|---|---|---|
| | **4** | **8** | **16** | **32** |
| Male | 20.4 (24.7) | 8.1 (13.3) | 6.7 (16.3) | 6.7 (16.3) |
| Female | 21.6 (34.8) | 16.7 (25.8) | 8.3 (20.4) | 11.7 (20.4) |
| Summary | **21** | **12.4** | **7.5** | **9.2** |

**Tab. 8.** Comparison of the mean *ERAC* and its standard deviation (in parentheses) depending on the number of GMM mixtures.

| **Voice / *ERAC* [%]** | **$G_{CMtype}$ ($N_{MIX}$ = 16)** | | |
|---|---|---|---|
| | **Spherical** | **Diagonal** | **Full** |
| Male | 63.3 (29.4)) | 11.7 (28.5) | 6.7 (16.3) |
| Female | 76.7 (21.6) | 16.7 (25.8) | 8.3 (20.4) |
| Summary | **70** | **14.2** | **7.5** |

**Tab. 9.** Comparison of the mean ERAC and its standard deviation (in parentheses) depending on the type of covariance matrix.

| **Phase / CPU time [s]** | **Covariance matrix type** | | |
|---|---|---|---|
| | Spherical | Diagonal | Full |
| GMM Training | 1.9 | 2.1 | 2.8 |
| Classification | 1.1 | 1.3 | 2.2 |
| **Total time** | **3** | **4.1** | **4.3** |

**Tab. 10.** Comparison of the computational complexity for different types of covariance matrices in the artifact classification experiment; $N_{MIX}$ = 16.

the *ERAC* values in Tab. 7. Therefore, a detailed analysis is necessary for correct threshold setting for determination of the 6th class. Contrary to general expectations, the achieved mean *ERAC* values do not fall for $N_{MIX}$ > 16 – they may even rise as documented in Tab. 8. Considering the fact that the 'Spherical' covariance matrix yields practically even 100% error rate in the majority of the tested artifact classes (over 70% in total), it is not suitable for this classification task – see the values in Tab. 9. The full covariance matrix and 16 mixtures were finally applied in our experiments to obtain an acceptable computational complexity and good results of the *ERAC* parameter for both

male and female voices. As regards the influence of the type of the covariance matrix on the overall CPU time consumption, the maximum value was measured for the 'Full' type and the minimum one for the 'Spherical' one – compare the numerical results in Tab. 10.

# 4. Conclusions

From the main point of view, the task of finding the alternative to the standard listening tests was fulfilled. The proposed and tested artifact detection, localization, and classification methods are functional and produce the results comparable with those obtained manually. The determined time durations of the performed tests and the listeners' feed-back information document differences between male and female listeners and their different approach in execution of the evaluation task: the female evaluators try to do it more carefully than the male ones resulting in a paradox – their results are practically worse than those of the male evaluators. At this point the "subjectivity" of the used method is well-founded and it also supports our aim to find objective evaluation methods.

At present, only one male and one female voice are implemented in the tested Czech TTS system working with the USEL-based synthesis method [2, 6, 7]. Therefore, the speech features determined from the synthetized sentences are not actually gender-dependent (male/female voice), but speaker-dependent (according to the voices used for building of the TTS inventory). Collecting of the databases of speech features from the synthetized sentences with and without artifacts was very difficult and time consuming. Therefore, at present only a small number of sentences were processed for usage in the ANOVA artifact automatic detection experiment. The proper choice of the used speech features in the input feature vector is very important. However, the choice of the optimal feature set for the artifact detection is not universal – different feature sets had to be used for the male and female voices. Generally, the detection accuracy depends, first of all, on elimination of statistical "similarities" between clean/artifact groups and a group of features from the tested sentence. In the case of the developed GMM-based artifact localizer, the realized auxiliary analysis has shown a considerable impact of different number of mixtures on the localization precision. Next, the existence of a principal influence of the accurate setting of the ROI on the precision of the artifact type classification was covered. If the ROI with the artifact is set incorrectly, the output error rate will rapidly increase up to 100% making the whole artifact detection, localization, and classification system useless. The presented artifact detection system processes only one artifact in an analyzed sentence. Two or more artifacts in one sentence could be found by dividing the speech signal into two parts and independent artifact detection and localization in each part. This step could be repeated several times, however, limited by the minimum time duration of the processed speech signal necessary for proper ANOVA analysis [9].

There are two imperfections which should be remedied in the near future to increase performance and accuracy of the whole developed artifact detection, localization, and classification system. The first drawback lies in the fact that mistakes caused by inappropriate segment duration are not treated as a special group of artifacts although they represent a considerable part of errors. At present, they are included in the 6th group but it could be worth distinguishing between the wrong segment length and the incorrect element. The second drawback stems from a relatively limited speech corpus of 40 sentences. A larger database with a sufficient number of sentences must be built to integrate all recognized speech artifacts produced by the TTS system based on the USEL synthesis. Finally, the results of the computation complexity in Matlab environment indicate the need for some optimization and implementation in a higher programing language for the real-time processing.

# Acknowledgments

The work has been supported by the Scientific Grant Agency of the Slovak Academy of Sciences and the Ministry of Education, Science, Research, and Sports of the Slovak Republic (VEGA 2/0001/17, VEGA 1/0905/17) and the Czech Science Foundation (GA16-04420S).

# References

[1] 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), 1996. p. 373–376. DOI: 10.1109/ICASSP.1996.541110

[2] TIHELKA, D., GRŮBER, M., JŮZOVÁ, M. Experiments with one-class classifier as a predictor of spectral discontinuities in unit concatenation. In Ronzhin, A. et al. (eds.) *Speech and Computer (SPECOM). LNAI 9811*. Springer, 2016, p. 296–303. DOI: 10.1007/978-3-319-43958-7_35

[3] NARENDRA, N. P., RAO, K. S. Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis. *Applied Soft Computing,* 2013, vol. 13, no. 2, p. 773 to 781. DOI: 10.1016/j.asoc.2012.09.023

[4] VEPA, J., KING, S., TAYLOR, P. New objective distance measures for spectral discontinuities in concatenative speech synthesis. In *Proceedings of the IEEE Workshop on Speech Synthesis*. Santa Monica (CA, USA), 2002, p. 223–226. DOI: 10.1109/WSS.2002.1224414

[5] MATOUŠEK, J., TIHELKA, D. On the influence of the number of anomalous and normal examples in anomaly-based annotation errors detection. In Sojka, P. et al. (eds.) *Text, Speech, and Dialogue (TSD). LNAI 9924*. Springer, 2016, p. 326–334. DOI: 10.1007/978-3-319-45510-5_37

[6] LEGÁT, M., MATOUŠEK, J. Identifying concatenation discontinuities by hierarchical divisive clustering of pitch contours. In Habernal, I., Matoušek, V. (eds.) *Text, Speech, and*

*Dialogue (TSD). LNAI 6836.* Berlin: Springer, 2011, p. 171–178. DOI: 10.1007/978-3-642-23538-2_22

[7] TIHELKA, D., MATOUŠEK, J., KALA, J. Quality deterioration factors in unit selection speech synthesis. In Matoušek, V., Mautner, P. (eds.) *Text, Speech and Dialogue (TSD), LNAI 4629.* Berlin: Springer, 2007, p. 508–515. DOI: 10.1007/978-3-540-74628-7_66

[8] DEY, S., MOTLICEK, P., MADIKERI, S., et al. Template-matching for text-dependent speaker verification. *Speech Communication,* 2017, vol. 88, p. 96–105. DOI: 10.1016/j.specom.2017.01.009

[9] PŘIBIL, J., PŘIBILOVÁ, A., MATOUŠEK, J. Detection of artefacts in Czech synthetic speech based on ANOVA statistics. In *Proceedings of the 38th International Conference on Telecommunications and Signal Processing (TSP).* Prague (Czech Republic), 2015, p. 1–5. DOI: 10.1109/TSP.2015.7296377

[10] PŘIBIL, J., PŘIBILOVÁ, A., MATOUŠEK, J. Experiment with GMM-based artefact localization in Czech synthetic speech. In Král, P., Matoušek, V. (eds.) *Text, Speech and Dialogue (TSD). LNAI 9302,* Cham Heidelberg: Springer, 2015, p. 23–31. DOI: 10.1007/978-3-319-24033-6_3

[11] BOERSMA, P., WEENINK, D. *Praat: Doing Phonetics by Computer (Version 5.4.22).* [Computer Program] Retrieved 2015-10-08. Available at: http://www.fon.hum.uva.nl/praat.

[12] PARALIC, M., JARINA, R. Iterative unsupervised GMM training for speaker indexing. *Radioengineering*, 2007, vol. 16, no. 3, p. 138–144. ISSN: 1210-2512

[13] LI, M., LIU, L, CAI, W., et al. Generalized i-vector representation with phonetic tokenizations and tandem features for both text independent and text dependent speaker verification. *Journal of Signal Processing Systems for Signal, Image, and Video Technology*, 2016, vol. 82, p. 207–215. DOI: 10.1007/s11265-015-1019-z

[14] STANEK, M., SIGMUND, M. Finding the most uniform changes in vowel polygon caused by psychological stress. *Radioengineering*, 2015, vol. 24, no. 2, p. 604–609. DOI: 10.13164/re.2015.0604

[15] RODELLAR-BIARGE, V., PALACIOS-ALONSO, D., NIETO-LLUIS, V., et al. Towards the search of detection in speech-relevant features for stress. *Expert Systems,* 2015, vol. 32, no. 6, p. 710–718. DOI: 10.1111/exsy.12109

[16] MEKYSKA, J., JANOUSOVA, E., GOMEZ-VILDA, P., et al. Robust and complex approach of pathological speech signal analysis. *Neurocomputing,* 2015, vol. 167, p. 94–111. DOI: 10.1016/j.neucom.2015.02.085

[17] MATOUŠEK, J., TIHELKA, D., ROMPORTL, J. Current state of Czech text-to-speech system ARTIC. In Sojka, P. et al. (eds.) *Text, Speech, and Dialogue (TSD). LNAI 4188.* Springer, 2006, p. 439 to 446. DOI: 10.1007/11846406_55

[18] TIHELKA, D., KALA, J., MATOUŠEK, J. Enhancements of Viterbi search for fast unit selection synthesis. In *Proceedings of INTERSPEECH*. Makuhari (Japan), 2010, p. 174–177.

[19] KALA, J., MATOUŠEK, J. Very fast unit selection using Viterbi search with zero-concatenation-cost chains. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence (Italy), 2014, p. 2569–2573. DOI: 10.1109/ICASSP.2014.6854064

[20] PŘIBIL, J., PŘIBILOVÁ, A. Evaluation of influence of spectral and prosodic features on GMM classification of Czech and Slovak emotional speech. *EURASIP Journal on Audio, Speech, and Music Processing,* 2013, vol. 2013, no. 8, p. 1–22. DOI: 10.1186/1687-4722-2013-8

[21] NABNEY, I. T. *Netlab Pattern Analysis Toolbox, Release 3.3.* [Online] Accessed 2015-10-02. Available at: http://www.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/downloads

## About the Authors ...

**Jiří PŘIBIL** (corresponding author), was born in 1962 in Prague, Czechoslovakia. He received his M.Sc. degree in Computer Engineering in 1991 and his Ph.D. degree in Applied Electronics in 1998 from the Czech Technical University in Prague. At present, he is a senior scientist at the Dept. of Imaging Methods, Inst. of Measurement Science, Slovak Academy of Sciences in Bratislava. His research interests are signal and image processing, speech analysis and synthesis, and text-to-speech systems.

**Anna PŘIBILOVÁ** received her M.Sc. and Ph.D. degrees from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology (FEEIT SUT) in 1985 and 2002, respectively. Since 1992 she has been working as a university teacher at the Radioelectronics Dept., and in 2014 she has become an associate professor at the Inst. of Electronics and Photonics, FEEIT SUT in Bratislava. The main field of her research and teaching activities is audio and speech signal processing.

**Jindřich MATOUŠEK** received his M.Sc. and Ph.D. 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 Dept. 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.