International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:2, 2017

An Approach for Vocal Register Recognition Based on Spectral Analysis of Singing

Aleksandra Zysk, Pawel Badura

***Abstract*—**Recognizing and controlling vocal registers during singing is a difficult task for beginner vocalist. It requires among others identifying which part of natural resonators is being used when a sound propagates through the body. Thus, an application has been designed allowing for sound recording, automatic vocal register recognition (VRR), and a graphical user interface providing real-time visualization of the signal and recognition results. Six spectral features are determined for each time frame and passed to the support vector machine classifier yielding a binary decision on the head or chest register assignment of the segment. The classification training and testing data have been recorded by ten professional female singers (soprano, aged 19-29) performing sounds for both chest and head register. The classification accuracy exceeded 93% in each of various validation schemes. Apart from a hard two-class clustering, the support vector classifier returns also information on the distance between particular feature vector and the discrimination hyperplane in a feature space. Such an information reflects the level of certainty of the vocal register classification in a fuzzy way. Thus, the designed recognition and training application is able to assess and visualize the continuous trend in singing in a user-friendly graphical mode providing an easy way to control the vocal emission.

***Keywords*—**Classification, singing, spectral analysis, vocal emission, vocal register.

I. Introduction

VOCAL registers refer to specific ways of voice emission during singing, depending on particular vibratory patterns of the vocal folds generating voice [1], [2]. The body areas responsible for resonance phenomena accompanying voice generation determine two main types of registers: chest and head. Whistle, falsetto registers and a vocal fry are also listed in, yet they are used by extremely small group of vocalists. Therefore, two terms are often used in description of singing modes: chest voice and head voice [3]. The chest voice is produced by the resonations within infraglottic cavity, trachea, bronchi, and thorax. In this case vocal folds vibrates on their edges. The head voice is generated by the upper resonators: larynx, throat, oral cavity, nasal cavity, and paranasal sinuses, vocal folds vibrates throughout their length. The chest voice can be perceived as stronger and darker than the head voice, therefore it is mostly associated with low notes. High notes require smaller resonation area which corresponds to the head voice. The awareness of the appropriate use of vocal registers during singing is an important and difficult task for beginner vocalist.

Lack of appropriate skills in operating vocal registers may cause serious phoniatric complications, such as vocal cord nodules, regurgitation or even loss of voice [4].

A. Zysk and P. Badura are with the Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta 40, 41-800 Zabrze, Poland (e-mail: olazysk94@gmail.com).

The application for vocal register recognition could be helpful for students who are training proper voice emission. According to modern trends in this area, mixing both head and chest register in all range of voice could be profitable during the process, since it provides healthy, natural position of the larynx and gives nice sounding timbre [5].

To the best of our knowledge, only few attempts have been done to address the vocal register recognition thus far from a highly specified point of view. Mysore et al. propose the mel-frequency cepstrum coefficients (MFCC) and support vector classification based approach for falsetto detection and report 95% accuracy over a set of 42 modal and falsetto notes [6]. Ishi et al. investigate the local power peaks and check their periodicity properties to detect the vocal fry singing [7]. The authors obtain a 74% detection accuracy and a 13% insertion error rate.

In this study, we present a framework for singing training support based on vocal register analysis. The automatic vocal register recognition (VRR) workflow covers sound acquisition, preprocessing, spectral analysis producing specific feature vector for each time frame, machine learning-based classification, and audio/graphical presentation of the results. The feature vector contains six spectral features: two different energy ratios, harmonicity, spectrum centroid, spread, and kurtosis. The support vector machine (SVM) has been selected as a classifier due to its recognized capability to handle two-class clustering issues and the ability to assess the input data in a fuzzy way. During the pilot study we involved ten professional female soprano singers able to control their voice emission. Their recordings of both chest and head voice have been used in initial spectral analysis, feature vector definition, classifier training, and validation. The classification accuracy has been validated in a set of experiments involving K-fold and leave-one-out cross validation schemes. The design and evaluation of the recognition framework have been followed by the implementation of an application with dedicated graphical user interface (GUI) for vocal register assessment support.

The paper is organized as follows. After the introduction in Section I, the materials and methods are presented in Section II in terms of acquisition setup, research sample, detailed data processing workflow, and GUI layout. Section III presents the experimental results along with the specification of employed validation schemes and discussion. The paper is concluded in Section IV.

II. Materials and Methods

A. Materials

The data were collected in repeatable conditions from ten professional women vocalists singing soprano, aged 19-29. During data acquisition subjects were standing 1 m from the microphone placed 80 cm above the ground. We used the Olympus WS-812 dictaphone with sampling frequency at 44100 Hz. Pitch was not imposed, all women were singing in such a way as to bring both vocal registers, according to their individual voice features. Data used for classifier training is ten pure tones lasting 0.309 s extracted from recordings of all subjects, for the chest and head voice, respectively.

B. Data Preprocessing

The audio signal is subjected to preprocessing procedures. First, the mean value (constant component) is subtracted from the signal. Then, the pre-emphasis filter is employed to enhance higher frequency components [8]. Then, the analysis moves into the frequency domain by means of a short-time Fourier transform (STFT) [8]. The STFT is calculated with 30 ms $\times$ 1 Hz time-frequency cells and a 0.2 ms time overlap. Due to the characteristics of a human voice, the further spectrogram analysis takes only its part between 100 Hz and 10 kHz.

Two procedures are employed over the spectrum amplitude S in a logarithmic scale in order to eliminate time frames disturbing the analysis: (1) the quiet periods and (2) transition frames. The silence frames are detected by means of a power spectral density (PSD) analysis. A total energy is estimated for the n^{th} frame as a sum of PSD samples:

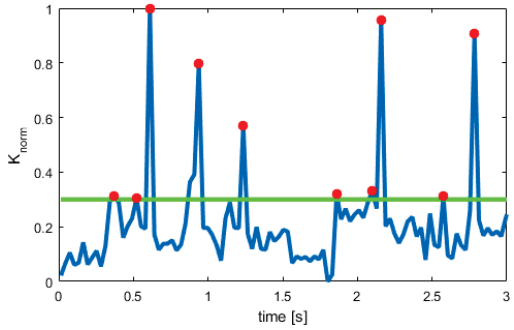
$$E(n) = \sum_{j=1}^{N_f} PSD(n,j), \tag{1}$$

where N_f denotes the number of frequency cells within the PSD. The frames with a $E(n)$ value below a constant threshold $E_{silence}$ are discarded and not subjected to feature extraction nor register recognition. The $E_{silence}$ value has been set to a 0.0001 of a maximum PSD value within the entire recording. Frames related to tone transitions are also removed due to the uncertainty they contribute to the recognition process. The transition detection procedure involves kurtosis of the spectrum, determined for the n^{th} frame as:

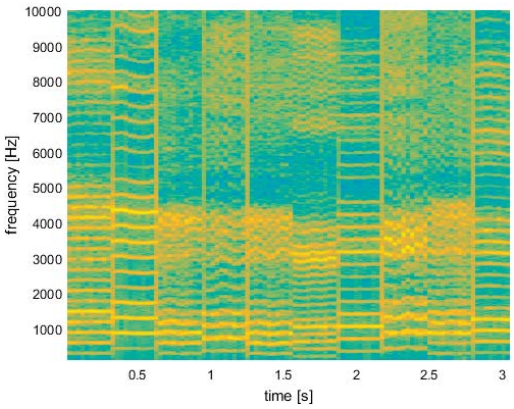
$$K(n) = \frac{\frac{1}{N}\sum_{j=1}^{N_f}\left(S(n,j) - \overline{S(n)}\right)^4}{\left(\frac{1}{N}\sum_{j=1}^{N_f}\left(S(n,j) - \overline{S(n)}\right)^2\right)^2}, \tag{2}$$

where $\overline{S(n)}$ is the average spectrum amplitude within the n^{th} frame. Then, the $K(n)$ sequence is normalized to a $[0,1]$ range. Finally, all local maxima of $K(n)$ exceeding the 0.3 value indicate transition frames excluded from further analysis (Fig. 1).

Before the feature extraction stage the dB spectrogram is normalized into a $[0,1]$ range.



(a)



(b)

Fig. 1 Illustration of a kurtosis-based detection of tone transition frames: the normalized $K(n)$ sequence with local maxima exceeding a threshold value (green line) indicated with red dots (a) and a corresponding spectrogram (b)

C. Feature Extraction

Six spectral features are determined for each frame:

1) Harmonicity H, describing the contribution of harmonics in the spectrum, calculated as a maximum value of the signal autocorrelation AC:

$$AC(n,k) = \frac{1}{N_f - 1} \cdot \sum_{j=1}^{N_f-k} \left(S(n,j) - \overline{S(n)}\right) \cdot \cdot \left(S(n,j+k) - \overline{S(n)}\right), \tag{3}$$

excluding zero delay ($k = 0$).

2) The spectral power distribution (SPD) for each frame is calculated and summed up in the range of 2000-6000 Hz (SPD_{2k}^{6k}) and 100-2000 Hz (SPD_{100}^{2k}). The ratio r_1^{SPD}:

$$r_1^{SPD}(n) = \frac{SPD_{2k}^{6k}}{SPD_{100}^{2k}} \tag{4}$$

stands for the second feature.

3) Similarly, the r_2^{SPD} ratio is determined for the SPD accumulated within ranges 1000-6000 Hz (SPD_{1k}^{6k}) and 100-1000 Hz (SPD_{1k}^{6k}):

$$r_2^{SPD}(n) = \frac{SPD_{1k}^{6k}}{SPD_{100}^{1k}} \tag{5}$$

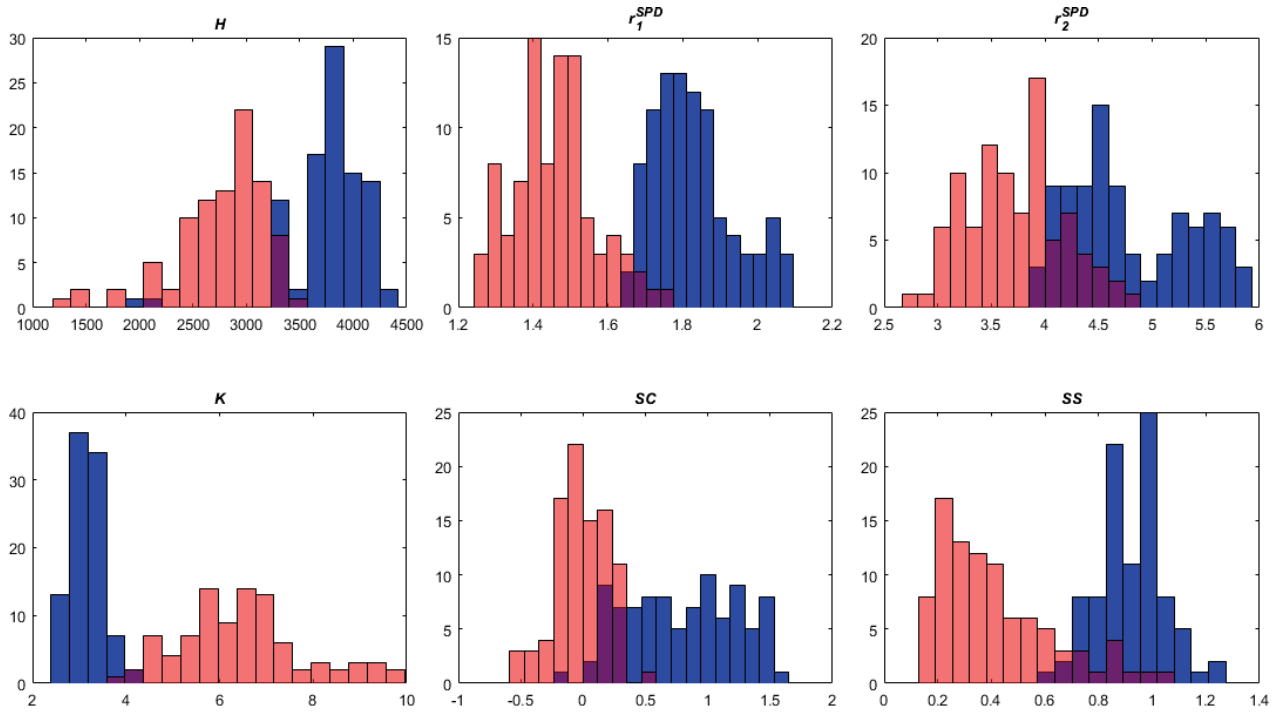


Fig. 2 Distributions of features throughout chest (blue) and head (red) registers

4) Kurtosis K is used as the other descriptor determined according to (2), yet over the normalised spectrogram.
5) The spectrum centroid SC known from the MPEG–7 standard [9] shows whether the spectrum is dominated by high or low frequencies. It is determined as:

$$SC(n) = \frac{\sum_{j=1}^{N_f} \left(\log_2 \left(\frac{F(j)}{1000} \right) \cdot PSD(n,j) \right)}{\sum_{j=1}^{N_f} PSD(n,j)} \quad (6)$$

where $F(j)$ [Hz] denotes the centre frequency of the j^{th} cell of the spectrum within a considered range (100-10000 Hz).
6) The MPEG–7 spectrum spread (SS) indicates a measure of focus around a centroid of the spectrum:

$$SS(n) = = \sqrt{\frac{\sum_{j=1}^{N_f} \left(\left(\log_2 \left(\frac{F(j)}{1000} \right) - SC(n) \right)^2 \cdot PSD(n,j) \right)}{\sum_{j=1}^{N_f} PSD(n,j)}}. \quad (7)$$

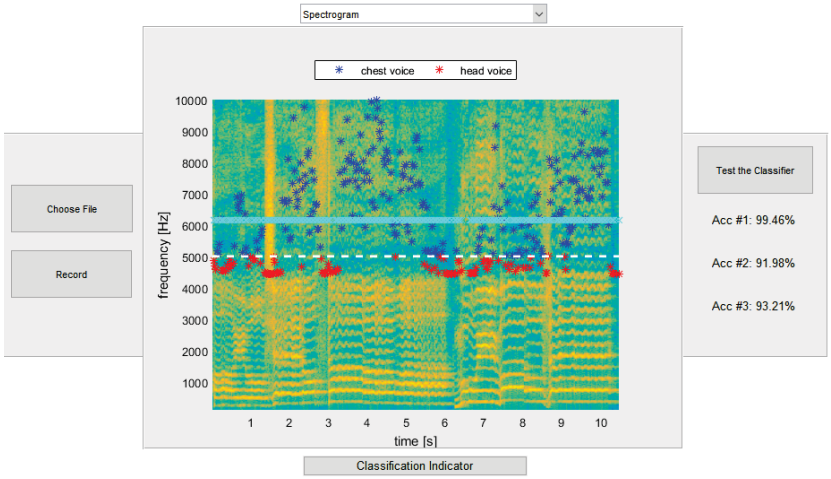
All features are normalised using minimum and maximum values predefined for each descriptor. Fig. 2 shows distributions of feature values over both chest and head voice frames from training data.

D. Classification

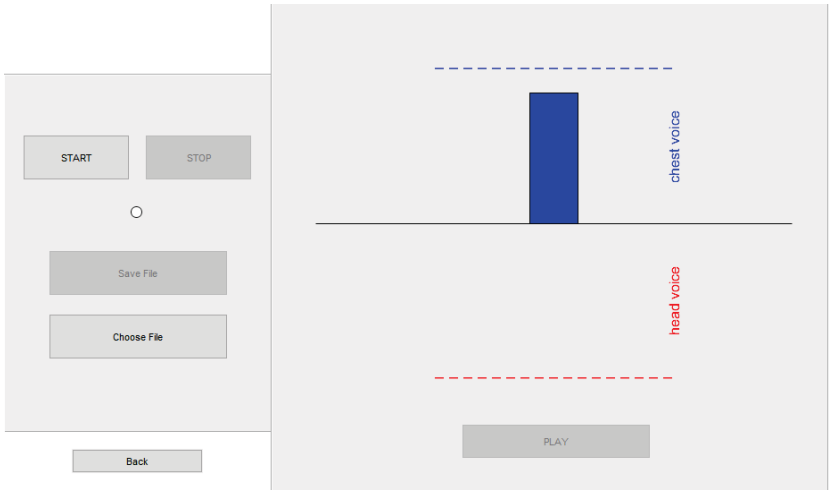
The two-class support vector machine classifier has been used as a classification tool [10] with input data subjected to a Gaussian radial basis mapping into kernel space. The trained classifier yields a binary clustering decision as a response to the input vector. Additionally, we use the information on the input vector distance d to the SVM discrimination hyperplane in the feature space. Such information can be used to assess the input frame in some sort of a fuzzy way.

E. Application for Vocal Register Recognition

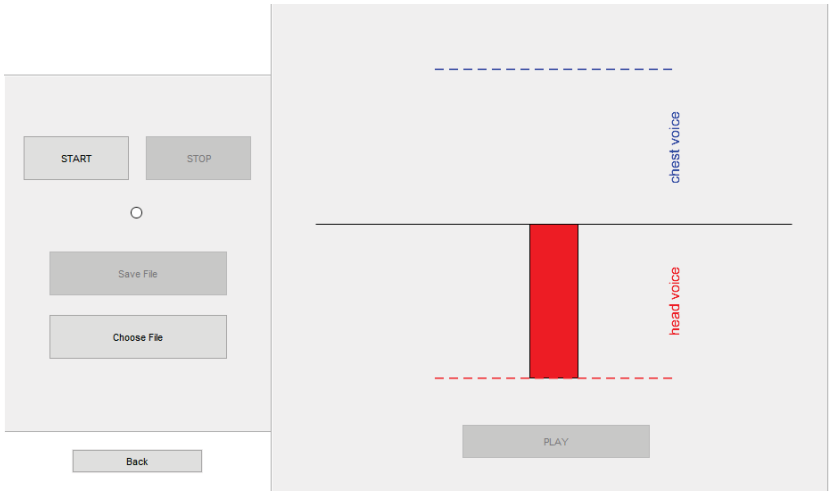
The VRR application allows the user to record her/his own audio signal with any sampling frequency. Both classification modes are employed by the application graphical user interface to visualize the register recognition results. On one hand, the entire spectrogram related to the recording under consideration is marked with either chest or head register indicator (Fig. 3a). The red marker indicates the head register; its vertical position on the bottom of the spectrogram shows the best fitting and in the middle of the spectrogram (white dotted line) — the weakest fitting. The blue marker corresponds to the chest register and, similarly, the greater the distance from the white line is, the better fitting it indicates. Markers located on the white line can be considered as mixed register. The overall classification indicator value averaged throughout all frames can also be shown, drawn as a cyan line in Fig. 3a. On the other hand, each acquired recording can be listened with line indication of a current sound position in the spectrogram and with a fuzzy visualization of a confidence level for chest or head register assignment using a bipolar bar or a dial gauge (Fig. 3b).



(a)

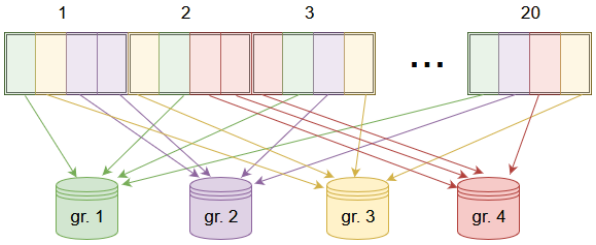


(b)

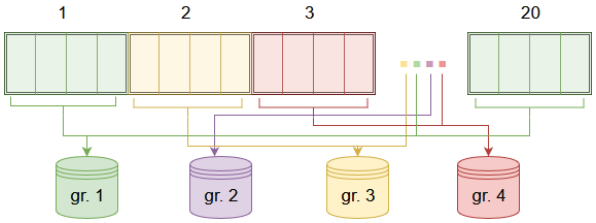


(c)

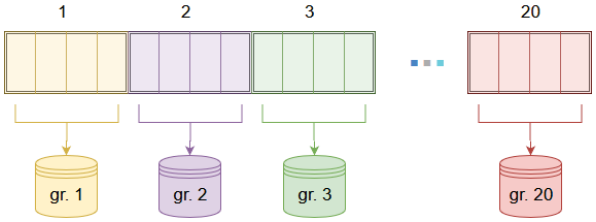
Fig. 3 Two modes for GUI visualization of vocal register recognition: on-spectrogram presentation (a) and real-time vocal register indicator (b)-(c). Blue and red colors refer to the chest and head registers, respectively. Distance from the middle line in (a) and absolute bar height in (b)-(c) reflect the recognition confidence level (SVM distance d)



(a) EXPERIMENT #1



(b) EXPERIMENT #2



(c) EXPERIMENT #3

Fig. 4 Illustration of cross validation schemes for vocal register classification assessment

III. RESULTS AND DISCUSSION

A. Experimental Results

The classification accuracy has been validated in three separate experiments:

- EXPERIMENT #1: all N time frames ($N = 114$) extracted from 20 pure tones (10 head and 10 chest register) throughout all reference signals recorded by professional singers have been subjected to a K-fold cross validation [11], with $K = 4$ (Fig. 4a).
- EXPERIMENT #2: all 20 pure tones have been divided into $K = 4$ groups for K-fold cross validation performed on their frames. The presence of at least one chest and head voice sample per each group was secured during the pseudorandom division of tones (Fig. 4b).
- EXPERIMENT #3: a leave-one-tone-out cross validation [11] has been performed over all 20 pure tones (Fig. 4c).

Each experiment is assessed on the basis of the accuracy metrics:

$$Acc = \frac{N_{ch} + N_h}{N} \cdot 100\%, \tag{8}$$

where N_{ch}, N_h denote the number of correctly recognized chest and head voice frames, respectively. In order to secure high level of reliability, each experiment was repeated 100 times. Obtained accuracies averaged throughout experiments are shown in Table I.

TABLE I
VOCAL REGISTER CLASSIFICATION ACCURACY SUMMARY

Experiment	Description	*Acc* [%]
EXPERIMENT #1	Frame-wise 4-fold cross validation	99.3 ± 0.4
EXPERIMENT #2	Tone-wise 4-fold cross validation	93.9 ± 1.3
EXPERIMENT #3	Leave-one-tone-out cross validation	93.2 ± 0.0

B. Discussion

The VRR system has proven its efficiency in vocal register recognition in a dedicated validation scheme. Repeatable accuracy metrics over 93% has been obtained over a reference database with annotated audio signals. The 99% accuracy in the first validation experiment can be explained by the presence of similar feature vectors in both training and testing datasets (there are multiple frames within each single tone distributed over different K-fold groups). However, high accuracies yielded by two latter experiments indicate high level of reliability of classification driven by selected spectral features. There is still room for the employment of a larger number of reference recordings in order to improve the VRR reliability and generalization capabilities.

The classifier trained with corresponding reference data has been established as an artificial expert within the VRR processing workflow. Supported by the graphical user interface and audio recording capabilities, the system can be employed as a tool for voice analysis and assessment. Fig. 5 shows some example visualizations for the assessment of audio signals related to some more or less complex melodies sung by the system users using certain vocal registers.

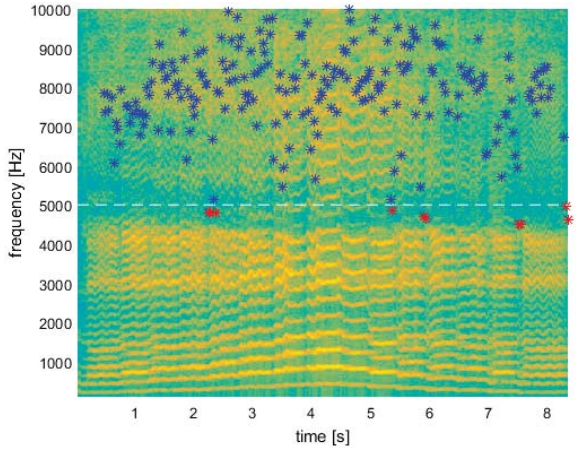
During the testing of the VRR with raw recordings we have noticed its generally correct work in case of professionally defined register types. Such observations have also been found during the analysis over a group of subjects less experienced in vocal emission. Main problems with recognition reliability have been identified in the following cases:

- transitions between tones and – obviously – silence periods make recognition unreliable; for that purpose two procedures have been employed at the preprocessing stage to eliminate such frames from the analysis;
- the system produces less consistent decisions while handling non-classical, improvisational vocals (e.g. jazz);
- sibilants and other noisy phonemes with high energy in the entire spectrum should not be considered register-distinguishable (note the "st" phoneme between 8^{th} and 9^{th} second in Fig. 5c); the recognition should be based mostly on vowel assessment.

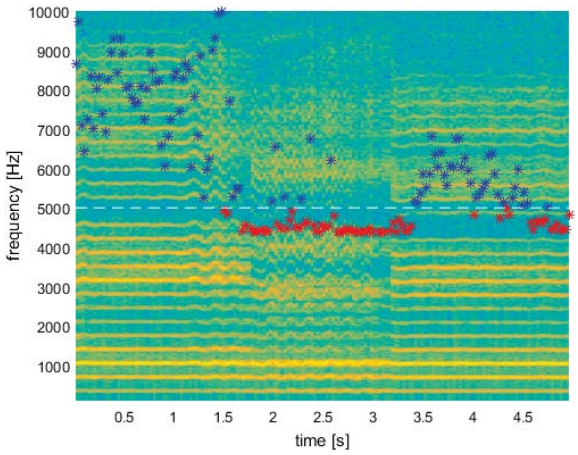
For the above reasons we employed the fuzzy mode for the register recognition and decided to rely on the graphical presentation of classification results. Moreover, some sort of low-pass or nonlinear filtering could be considered over the classification result signal in order to avoid artefacts.

IV. CONCLUSION

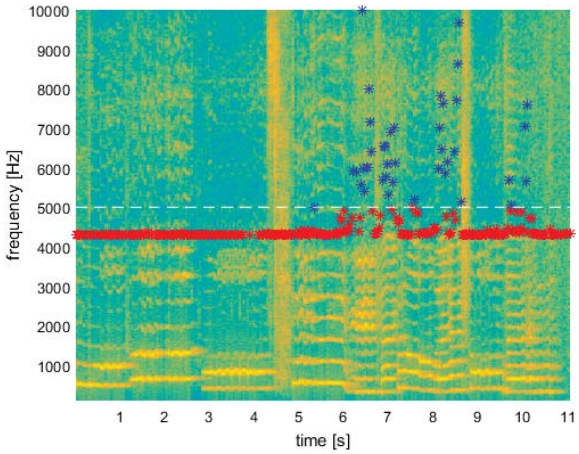
The accurate VRR classification proves essential differences in spectral distribution of singing signal for both considered vocal registers, whereas the graphical user interface provides user-friendly presentation of the singing analysis results. This



(a) The A major scale in chest register



(b) Tone f^1 in three different registers: chest, head, and mixed



(c) Part of Franz Schubert's *An die Musik*, head register

Fig. 5 Examples of vocal recognition results

conclusion leads to possibility of supporting voice emission lessons and individual practicing with computer application, especially at the beginning of training, when students are not able to discern objectively which natural resonators they are using while singing different sounds.

References

[1] J. Large, "Towards an integrated physiologic-acoustic theory of vocal registers," *The NATS Bulletin*, vol. 28, pp. 30–35, 1972.
[2] R. L. Whitehead, D. E. Metz, and B. H. Whitehead, "Vibratory patterns of the vocal folds during pulse register phonation," *The Journal of the Acoustical Society of America*, vol. 75, no. 4, pp. 1293–1297, Apr. 1984.
[3] J. Stark, *Bel Canto: A History of Vocal Pedagogy*. University of Toronto Press, 2003.
[4] R. H. Colton, J. K. Casper, and R. Leonard, *Understanding Voice Probems: A Physiological Perspective for Diagnosis and Treatment*. Lippincott Williams & Wilkins, 2006.
[5] A. Frisell, *The Tenor voice: a personal guide to acquring a superior singing technique*. Branden Publishing Company, 2007.
[6] G. J. Mysore, R. J. Cassidy, and J. O. Smith, "Singer-dependent falsetto detection for live vocal processing based on support vector classification," in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*. Institute of Electrical and Electronics Engineers (IEEE), 2006.
[7] C. T. Ishi, K.-I. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 47–56, Jan. 2008.
[8] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing (2^{nd} Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999.
[9] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley & Sons, 2002.
[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[11] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.