Evaluation of Singer's Voice Quality by Means of Visual Pattern Recognition

Paweł Forczmański, *Szczecin, Poland*

Summary: The article presents a description of the algorithm of singing voice quality assessment that uses selected methods from the field of digital image processing and recognition. It adopts the assumption that an audio signal with recorded vocal exercise can be converted into a visual representation, and processed further, as an image. Presented approach is based on generating a sound spectrogram of a sample in the form of a rectangular matrix, objective improvement of its visual quality based on local changes in brightness and contrast, and scaling to a fixed size. Then, it uses a two-step approach: the construction of a representative database of reference samples and the identification of test samples. The process of building the database uses two-dimensional linear discriminant analysis. Then, the recognition operation is carried out in a reduced feature space that has been obtained by two-dimensional Karhunen-Loeve projection. Classification is done by a variant of Support Vector Machines approach. As it is shown, the results are very encouraging and are competitive to the most powerful state-of-the-art methods.

Key Words: Singing quality–Image recognition–Image processing–Spectrogram–Short-time Fourier transform–Linear discriminant analysis–Support vector machine.

INTRODUCTION

Problem definition

Analysis of human voice is one of the most interesting tasks of multimedia systems. Within this problem, we can distinguish three main research directions: speech recognition (in terms of content), speaker recognition (identification), and evaluation of speech quality. Although the identification of persons on the basis of the registered voice and speech recognition in terms of content are tasks fairly well described in the literature and implemented in practice, the evaluation of voice quality is a problem that is still not fully solved. It may involve the automatic evaluation of voice quality, for example, in the process of language learning and assessment of the degree of training of lectors and singers. It is worth noting that the sensitivity, which characterizes human sense of hearing, has long been achievable for technical equipment—each of the physical quantities characterizing the speech signal can now be specified much more precisely using computerized analyzers than using the sense of hearing. At the same time, however, man is still able to use the acquired information from the voice signal in a more effective manner. This is due to the fact that the sense of hearing and the human nervous system are highly specialized and adapted through evolution to collect and analyze speech signal; however, occurring processes are not fully understood. This makes serious difficulties in implementation of computer algorithms aimed at such problems. Voice analysis is the subject of research of specialists in many fields: phoneticians, phoniatrists, speech therapists, and specialists in telecommunications, but in spite of many studies, speech signal must be considered complex and difficult to complete interpretation (ie, comparable to the analysis performed by human). The speech signal contains a complex information that helps to receive the basic meaning of speech and gives an opportunity to detect additional features, such as interlocutor's sex, age, health status, mood, education, and others. Analysis of the literature shows that speech signal can be described by a number of numerical parameters. It should be noted that there are various known techniques aimed, for example, at measurements of voice acoustic phonetics, medicine, and automatic speech recognition.[1] Thus, it seems that there is no need to seek for new methods of representation of speech signal, but the focus should be on the selection and use of existing ones.

Singing and speech signal are extremely complex, when it comes to formal description. The concept of singing is strongly related to the concept of quality. In the traditional approach, singing is usually assessed by an expert or group of experts related to the issue of voice, and this analysis is perceptual.[2] Because this process is human-centric and highly subjective, it may be interesting to provide some measures and techniques to make it more objective.

It should be noted that automation of the singing quality evaluation process can have multiple purposes, including supporting the learning process of singing and vocal skills with singers, supporting the classification of singers in terms of vocal advancement and suitability for specific vocal assignments, together with the identification of disorders in emission and potential health problems. It is easy to imagine that a computer system performs an automatic evaluation of singer's advancement by means of analyzing a small vocal exercise recorded using a simple microphone. This examination can be used in choir rehearsal or as a routine check of singer's physical shape.

It is therefore important to find a subset of parameters measured for sound for which the interindividual variability is significantly smaller than the variation resulting from the level of vocal advancement. In this work, it is also assumed that previously described process can be performed by means of timing and frequency-subcontours for sound.

Therefore, this work focuses on the task of singing quality evaluation in the context of automated classification of the level

Accepted for publication March 2, 2015.

From the Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, Szczecin, 52 Zolnierska St., 71-210 Szczecin, Poland.

Address correspondence and reprint requests to Paweł Forczmański, Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, Szczecin, 52 Zolnierska St., 71-210 Szczecin, Poland. E-mail: pforczmanski@wi.zut.edu.pl

Journal of Voice, Vol. 30, No. 1, pp. 127.e21-127.e30
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2015.03.001

of singer's training. The proposed process uses objective features taken from voice signal and can be applied in many practical situations, for example, when evaluating singer's abilities and personal vocal skills or to detect potential anomalies in voice production.

The article is organized as follows. The rest of introductory part presents some related works. The second section provides a description of the developed algorithm as well, as the benchmark data set consisting of real vocal samples collected from choir singers. The third section presents numerical experiments and discusses their results. The last section concludes the article.

Related works

There is a large number of approaches related to the automatic evaluation of singing quality, which can be found in the literature. One of the proposed measures of singing quality is singing power ratio (SPR).[3] It is calculated in the spectral domain and is defined as the ratio between the highest peak amplitude for frequencies from range 2–4 kHz and the highest peak in the range of 0–2 kHz. According to Omori et al,[3] SPR can be used to distinguish between voice samples with extremely varying advancement levels. The same feature has been used together with a set of other factors in another study by Nakano et al,[4] to determine the differences between training of singing students. Experiments conducted on 55 people show that it is not possible to clearly distinguish subjects. Taking into account the results obtained, it can be concluded that SPR can be used to distinguish singers only if their skills vary in a significant way.

Another popular criterion for assessing the singing quality is intonation accuracy used by Murry[5] in terms of an ability to perform a pitch-matching task. It is considered as one of the measures, which is independent of individual characteristics and melodic features.[6] Intonation accuracy is often used in combination with other features for a more reliable assessment of singing quality. In the studies by Kostek and Żwan,[7,8] they showed how to evaluate singing quality on the basis of only one sound (a vowel). Described classification of singing voice has been carried out on two levels: on the basis of assessment of voice itself (amateur and semiprofessional) and its type (bass, baritone, tenor, alto, mezzo-soprano, and soprano). A set of parameters describing a song was used in this solution for the construction of a feature vector, in which a single vowel for each singer was evaluated by six experts. The resulting quality index calculated for 2690 recordings was used for training an artificial neural network. Obtained classification accuracy reached 84–90%, depending on the features used. In the study by Jha and Rao,[9] they presented a similar approach, which uses an analysis of vowels to assess the quality of singing voice. In this solution, two signal characteristics were used, namely an envelope of the spectrum and a pitch. They have been subjected to classification by means of Gaussian mixture models and linear regression. The accuracy achieved 76–89%. Neural networks used by Hariharan et al for operations on speech signal in time domain were shown in the study by Hariharan et al,[10] with over 98% accuracy of classification in case of normal and pathologic conditions. The classification of speech signals in reduced feature space obtained by principal component analysis was presented.[11–13] Another approach to classification using linear discriminant analysis (LDA) was presented in the study by Lee et al.[14] The tests were performed on samples of normal and pathologic voice and shown 83% accuracy of the proposed method.

As it has been shown in the scientific literature from the field of signal processing, most of the methods use physical characteristics of sound through low-level feature vectors consisting of a set of coefficients calculated in the time domain, frequency, or cepstrum. Their recognition and classification are based primarily on a one-dimensional approach to data. Resulting effectiveness varies and depends on the initial conditions and problem nature. In case of classification of singing quality, presented methods achieve the accuracy of about 90%.

In contrast, this article focuses on a two-dimensional approach to singing signal presented in a form of a matrix. Because sound signal is a one-dimensional function (eg, amplitude), it is not easy to capture all its variability over time. When we add another dimension to this representation, we can depict sound as matrix and store much more sound features in one, compact structure. The typical representation of any two-dimensional matrix, in computer science, is an image; therefore, it seems to be natural to use algorithms from the field of digital image processing and recognition for signal processing. Summarizing, the proposed approach is based on the observation that the visual, two-dimensional representation of the signal of singing voice carries much more information that in case of standard and limited vector representation. By using selected set of methods aimed at image processing, it is assumed to obtain higher accuracy of singer's voice quality estimation in comparison with established methods.

MATERIALS AND METHODS

Initial assumptions

Numerical analysis of human voice must take into account its time-frequency structure because it depends on the phenomena encountered with the production of acoustic signal. Details of the anatomy of vocal tract (geometric dimensions, acoustic impedance of tissues) are different for each person, and each difference is reflected in the parameters of produced acoustic voice. On the other hand, it can be assumed that certain unique features are responsible for the quality of the produced voice and indirectly for the level of singer's training. These features may be also common for larger groups of singers. Such formulated assumptions allow to build a hierarchical database in which a single class will consist grouped voice samples in terms of similar degree of training. Previously described features may be represented in a graphical form which can capture their variability in a more adequate way, as opposed to strictly numerical, one-dimensional parameters.

Processing outline

An image of voice spectrum (so-called spectrogram) is a well-known method of sound representation. When used as a result