ARCHIVES OF ACOUSTICS
33, 1, 65–71 (2008)

AUTOMATIC SINGING QUALITY RECOGNITION EMPLOYING ARTIFICIAL NEURAL NETWORKS

Paweł ŻWAN

Gdańsk University of Technology
Multimedia Systems Department
Narutowicza 11/12, 80-952 Gdańsk, Poland
e-mail: zwan@sound.eti.pg.gda.pl

(*received June 15, 2007; accepted November 30, 2007*)

The aim of the paper is to determine how quality of a singing voice can be recognized automatically. For this purpose, a database of singing voice sounds with samples of voices of trained and untrained singers was created and is presented. The methods of a singing voice parameterization are shortly reviewed and a set of descriptors is outlined. Each of the presented samples is parameterized and judged by experts, and the resulting feature vectors and quality scores are then used to train an artificial neural network. A comparison between experts' judgments and automatic recognition results is performed. Finally, statistical methods are applied to prove that an artificial neural network is able to automatically determine the quality of a singing voice with the accuracy very similar to expert assessments. The paper includes the discussion of results and presents derived conclusions.

Keywords: singing voice, neural networks, automatic sound recognition.

1. Introduction

Automatic sound recognition requires the process of feature extraction. Parameterization algorithms of musical signals are well developed and discussed in the domain of music information retrieval (MIR) and automatic speech recognition [3, 7]. Some MIR systems allow to index media content automatically. They are based on the definitions of parameters describing timbre differences between musical instruments. On the other hand, automatic speech recognition systems perform automatic text recognition with very good results. Singing voice is the domain where those two fields meet, because it is produced by the same vocal organs as speech and is considered a musical instrument by musicologists. In this aspect, some speech parameters can be used. However, due to the artistic and instrumental character of singing, they need to be modified and complemented by the designed new ones.

This is particularly important when the quality factor is concerned. The quality factor is defined as a subjective measure of whether the voice belongs to an amateur or to

a professional singer. In some cases, the classification into such two quality groups is very difficult, and this is mostly due to the subjectiveness of expert judgments. Since no quality class definition is presented in literature, a novel approach to quality description must be introduced and then statistically tested. In the paper, a comparison between the results of automatic classification and the plots of experts' judgments distribution is done, and it is shown that intelligent decision systems trained on the basis of professional judgments perform very similarly to experts and are able to determine the quality of singing voice with a very similar precision.

2. Singing voice database

A database of singing voices was set up, that stored 1440 sound samples recorded by 42 different vocalists. Each of the vocalists recorded 5 vowels: "a", "e", "i", "o", "u" at several sound pitches belonging to their natural voice scales. Sound files were recorded in a studio with a sampling frequency of 44.1 kHz and a 16-bit resolution using the Neumann TLM 103 microphone. There were three groups of vocalists: amateurs (Gdańsk University of Technology Choir vocalists), semi-professionals (Gdańsk Academy of Music, Vocal faculty students) and professionals (qualified vocalists, graduates of the Vocal Faculty, of the Gdańsk Academy of Music). The quality of each of the recorded vowels was judged by experts, who were singing teachers and professional vocalists, by assigning scores (1, 2, 3, 4, 5) to sound samples. Since it was not possible to avoid half-point scores (1; 1.5; 2; 2.5; 3; 3.5; 4; 4.5; 5), finally each of the singing voice sound samples was qualified to one of 9 quality classes (QC1 – QC9). All experts were also checked as to the stability of their evaluation scores, therefore the average judgment of all 6 experts was assumed to be the final sound quality score.

In order to verify how precise the experts were in assessing voice samples, a distribution of their judgments was calculated as a function of a quality class (the quality class was assumed as an average judgment of all 6 experts). The function values were the differences (..., -3,-2,-1, 0, 1, 2, 3, ...) between the scores of a single expert and an average score of all experts; its value was 0 if the judgment was equal. The calculations were done for each of the 1440 analyzed voice samples and for each expert separately. This resulted in 6 expert plots, showing how precise the experts were in judging voices in the function of a judged class. In Fig. 1, expert precision curves averaged for 6 analyzed experts are presented individually for each quality class.

The precision of the experts' judgment was not equal for all category classes. The evaluations turned out to be less precise for middle quality classes (approx. 30% samples were judged as equal to the average judgment).

For quality classes QC1 and QC9, the ability to judge precisely was much better. Over 70% evaluation scores of a single expert were equal to the average judgment. More detailed information on the data from Fig. 1 are presented in Table 1, where columns denote quality classes, and rows indicate experts' judgments.

Table 1. Experts' judgments versus recognized quality classes.

[%]	QC1	QC2	QC3	QC4	QC5	QC6	QC7	QC8	QC9
C1	72.5	27.5	0	0	0	0	0	0	0
C2	35.5	38.3	18.8	7.5	0.2	0	0	0	0
C3	15.3	20.8	29.6	21.3	10	2.5	0	0	0
C4	3.8	6.7	27.5	29.6	19.5	12.5	0.4	0	0
C5	0	2.5	5.8	27.5	39.2	20.8	4.2	0	0
C6	0	0	0.3	12.6	19.7	32.4	29.9	5.1	0
C7	0	0	0	1.4	6	20	47.2	22.1	3.2
C8	0	0	0	0	0	5	22.1	43.8	29.2
C9	0	0	0	0	0	0.5	10.6	15	74

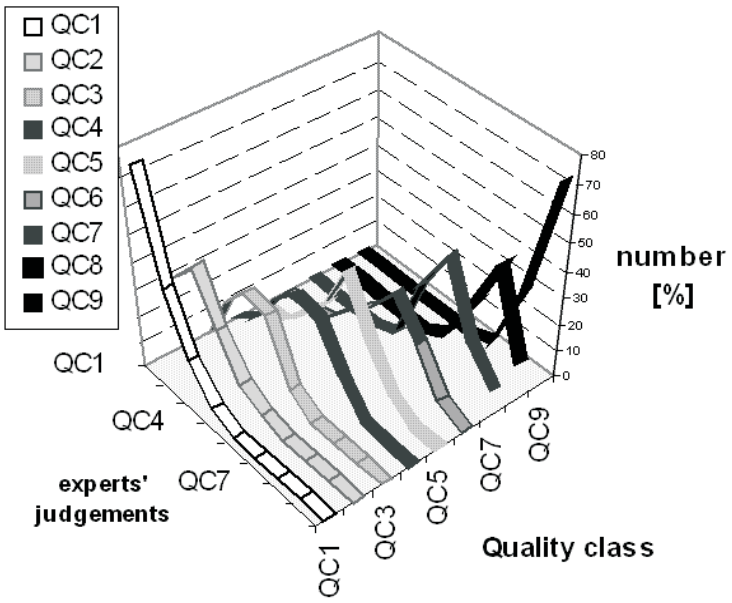

Fig. 1. Experts' "precision plots" in function of quality class.

3. Singing voice parameterization

Singing voice parameters were defined and described in detail in other publications of the author of this paper [11–13]. Since the paper focuses on the results of automatic recognition, only the parameters contained in feature vectors will be outlined.

Singing is produced by the vibration of vocal cords and resonances within the throat and head cavities that shape the timbre and power of an outgoing sound. The reso-

nances generate formants in the spectrum of produced sounds. Lower formants (middle frequency lower then 2 kHz) are related to articulation producing different vowels, higher formants (middle frequency higher than 2 kHz) characterize mainly timbre and voice type qualities[6, 8]. The formant of the middle frequency around 3.5 kHz is called "singer's formant" and its relation to voice quality is proved in a reach literature concerning singing [1, 9, 10]. However, the resulting timbre and power of an outgoing vocal sound is formed by the interaction between two factors, namely the glottal source and the vocal tract resonance characteristics. The relation between those factors is not simple but becomes less complicated when we assume the linearity of the vocal tract filter. Since there exists an analogy between the FIR filtering and a singing sound in the proposed model – which can be represented as a convolution of the glottal source and impulse response of the vocal tract – singing voice parameters can be divided into two main groups related to these two factors. In literature, some inverse filtration methods for deriving glottis parameters are presented, however, they are inefficient due to the phase shift problems of the vocal tract filter [10]. In this aspect only the parameters of the vocal tract formants can be calculated directly from the inverse filtering analysis since they are defined in the frequency domain. To calculate formant parameters the Warped Liner Prediction (WLPC analysis) was chosen [5].

Glottal source parameters, which are defined in the time domain, are not easy to compute from the inverse filtration but within the context of a singing voice quality, their stability in time rather that their objective values seems to be important. Time resolution of the analysis is crucial, single periods of a sound in a sonogram must be observed, thus the analysis with small frames and substantial overlapping seems to be the most proper approach. For each of the frequency bands, a sonogram consists of a set of n sequences $S_n(k)$, where n is the number of a frequency band and k is the number a sample. Since the aim of parameterization is to describe the stability of energy changes in sub-bands, the autocorrelation function in time of sequences $S_n(k)$ is employed. The more frequent and stable the energy changes in a sub-band are, the higher are the values of the maximum of the autocorrelation function (for index not equal to 0).

Another group of parameters are vibrato and intonation parameters [2, 10]. It is clear that a person who neither holds the pitch nor gets a stable vibrato cannot be judged as a good singer. In order to calculate vibrato parameters, the pitch contour needs to be extracted. There are several methods for an automatic sound pitch extraction, of which autocorrelation seems to be the most appropriate [11].

Parameterization of vibrato depth and frequency (f_{VIB}) may be insufficient to asses singing quality, thus other pitch contour parameters are introduced, such as "*periodicity"* of vibrato pitch contour, defined as the maximum value of the autocorrelation of the pitch contour function (for index not equal to 0), "*harmonicity"* of vibrato calculated as Spectrum Flatness Measure for the spectrum of the pitch contour; "*sinusoidality"* of vibrato VIB_S defined as the similarity of the parameterized pitch contour to the sine waveform [11, 12].

In order to complement dedicated parameters, some more general signal features such as descriptors of audio content contained in the MPEG-7 standard were used. Al-

though those parameters are not related to the singing voice biomechanics, they are useful in the process of a singing voice recognition [12].

4. Experiments

All 1440 singing voice sounds were parameterized using all 331 parameters. For the purpose of automatic singing voice recognition, a feed-forward neural network was employed. Singing voice sounds were divided into two sets for training and testing purposes. Since the aim of the network was to mimic experts' judgments, the network should have returned a value proportional to an expert's judgment (1–5). Thus, the output layer consisted of 1 neuron with a linear activation function. The output neuron value was quantized as presented in Table 2.

Table 2. Quantization of output neuron values into quality classes.

Output neuron values	Quality Class
$\langle -1, -0.78)$	QC1
$\langle -0.78, -0.56)$	QC2
$\langle -0.56, -0.34)$	QC3
$\langle -0.34, -0.12)$	QC4
$\langle -0.12, 0.11)$	QC5
$\langle 0.11, 0.34)$	QC6
$\langle 0.34, 0.56)$	QC7
$\langle 0.56, 0.78)$	QC8
$\langle 0.78, 1)$	QC9

Table 3 presents the results of automatic recognition. The network was tested using sound samples of singers whose vices were not used in the training phase. Presented values correspond to the average result of 42 trained networks, with the tested vocalist being changed for each network, a so-called k-fold cross validation method was employed [4].

The efficiency defined as the number of sounds recognized as an adequate quality score versus the total number of sounds seems to be low, although in order to analyze it precisely it should be compared to "expert precision plots" presented in Fig. 1. The number of quality classes was set arbitrary and the experts were not able to classify singing voices with such precision as (1–5) scores. While comparing Tables 1 and 3, it can be noticed that the automatic recognition performance and expert precision curves are very similar in distribution versus a quality class number (e.g. in Fig. 2 adequate curves for quality classes 6 and 8 are presented, their distribution is very similar).

The distribution of the automatic recognition error and experts' precision plots were compared statistically by using the Pearson's autocorrelation measure and the results were greater than a critical value of 0.834 (for $\alpha = 0.005$). This proves that differences between the results are non relevant statistically.

Table 3. Experts' judgments versus recognized quality classes.

[%]	QC1	QC2	QC3	QC4	QC5	QC6	QC7	QC8	QC9
QC1	60.0	30.0	10.0	0	0	0	0	0	0
QC2	16.1	33.0	29.5	13.8	6.3	1.3	0	0	0
QC3	6.2	19.9	39.8	18.1	9.7	4.9	0.9	0.4	0
QC4	0	11.1	11.1	35.6	22.2	13.3	6.7	0	0
QC5	0	0	4.8	17.3	38.5	32.7	5.8	1.0	0
QC6	0	0.8	1.6	7.8	15.5	37.2	24.8	9.3	3.1
QC7	0	0	0.4	1.8	5.3	15.9	34.4	29.1	13.2
QC8	0	0	0	0	3.7	9.2	29.5	42.8	14.8
QC9	0	0	0	0.6	0.6	4.4	11.9	37.7	44.7

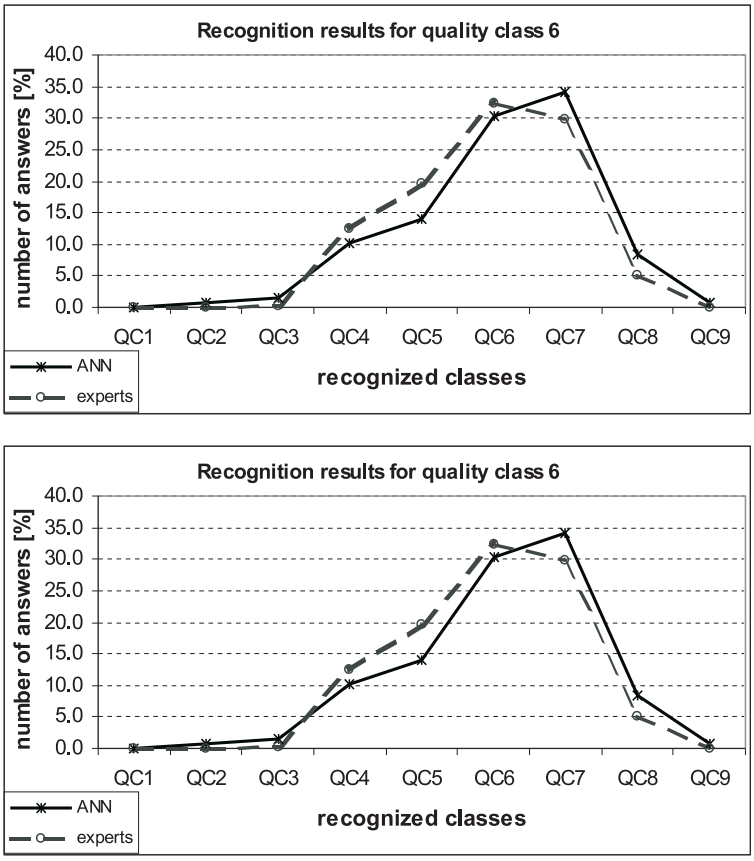

Fig. 2. Automatic recognition error distribution and expert precision curves for quality classes: QC6 and QC8.

5. Conclusions

The implemented system recognized singing quality with a precision very similar to experts judgments. The subjective factor was limited by selecting several judges and defining the voice quality score as their average judgment. By using intelligent decision methods, it was possible to train a feed-forward neural network based on experts' judgments. In order to check general qualities of the system, the network was tested by the samples of those singers whose voices were not used in the training phase.

An arbitrary number of quality classes is the reason why the efficiency of the automatic recognition system, defined as the number of sounds assigned to an adequate quality class, seems to be low. However, an analysis of the distribution of experts' judgments shows that the experts are not precise in classifying a singing voice quality and they have problems with assigning samples to quality classes. A comparison between the results of automatic recognition and the plots representing experts' judgments shows a strong correlation proven statistically. Therefore the automatic recognition system performance is similar to experts' judgments.

References

[1] BLOOTHOOF G., *The sound level of the singer's formant in professional singing*, J. Ac. Soc. Am., **79**, 2028–2032 (1986).

[2] DIAZ J. A., ROTHMAN H. B., *Acoustic parameters for determining the differences between good and poor vibrato in singing*, Proc. 17th International Congress on Acoustics, Rome, VIII, 110–111 (2001).

[3] DOWNIE S., *Music information retrieval, annual review of information science and technology*, **37**, 295–340 (2003).

[4] GOUTTE C., *Note on free launches and cross-validation*, Neural Computation, 2000.

[5] HARMA A., KARJALAINEN M., VALIMAKI V., LAINE U., *Frequency-warped signal processing for audio applications*, J. Audio Eng. Soc., 2000.

[6] JOLIVEAU E., SMITH J., WOLFE J., *Vocal tract resonances in singing: the soprano voice*, J. Acoust. Soc. Am., **116**, 2434–39 (2004).

[7] KOSTEK B., *Perception-based data processing in acoustics. applications to music information retrieval and psychophysiology of hearing*, Springer Verlag, Series on Cognitive Technologies, Berlin, Heidelberg, New York 2005.

[8] MILLER D. G., *Formant Tuning in a Professional Baritone*, Journal of Voice, **4**, 231–237 (1990).

[9] ROTHMAN H. B., *Why we don't like these singers*, Proc. 17th International Congress on Acoustics, **VIII**, 114–115 (2001).

[10] SUNDBERG J., *The science of the singing voice*, Northern Illinois University Press, Illinois 1987.

[11] ŻWAN P., *Expert system for automatic classification and quality assessment of singing voices*, Proc. 121 AES Convention, USA, San Francisco 2006.

[12] ŻWAN P., SZCZUKO P., KOSTEK B., CZYŻEWSKI A., *Automatic singing voice recognition employing neural networks and rough sets*, RSEISP, LNAI, Warszawa 2007.

[13] ŻWAN P., *The expert system for objectivization of singing voice judgements* [in Polish], Ph.D. Thesis, Gdańsk Univ. of Technology, Multimedia Systems Dept., 2007.