Seeking the SuperStar: automatic assessment of perceived singing quality

Johanna Bohm, Florian Eyben, Maximilian Schmitt, Harald Kosch, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Bohm, Johanna, Florian Eyben, Maximilian Schmitt, Harald Kosch, and Björn Schuller. 2017. “Seeking the SuperStar: automatic assessment of perceived singing quality.” In *International Joint Conference on Neural Networks (IJCNN), 14-19 May 2017, Anchorage, AK, USA*, 1560–69. Piscataway, NJ: IEEE. https://doi.org/10.1109/ijcnn.2017.7966037.

Seeking the SuperStar: Automatic Assessment of Perceived Singing Quality

Johanna Böhm[1], Florian Eyben[1,2], Maximilian Schmitt[3], Harald Kosch[4], Björn Schuller[1,2,3,5]
[1]Machine Intelligence & Signal Processing group, Technische Universität München, Germany
[2]audEERING GmbH, Gilching, Germany
[3]Chair of Complex and Intelligent Systems, University of Passau, Germany
[4]Chair of Distributed Information Systems, University of Passau, Germany
[5]Department of Computing, Imperial College London, U. K.

Abstract—**The quality of the singing voice is an important aspect of subjective, aesthetic perception of music. In this contribution, we propose a method to automatically assess perceived singing quality. We classify monophonic vocal recordings without accompaniment into one of three classes of singing quality. Unprocessed private and non-commercial recordings from a social media website are utilised. In addition to the user ratings given on the website, we let both subjects with and without a musical background annotate the samples. Building on musicological foundations, we define and extract acoustic parameters describing the quality of the sound, musical expression and intonation of the singing. Besides features which are already established in the field of Music Information Retrieval, such as *loudness* and *mel-frequency cepstral coefficients*, we propose and employ new types of features which are specific to intonation. For automatic classification by supervised machine learning methods, models predicting the subjective ratings and the user ratings on the social media website are learnt. We perform an exhaustive evaluation of both different classifiers and combinations of features. We show that the performance of automatic classification is close to that of human evaluators. Utilising support vector machines, an accuracy of classification of 55.4 %, based on the subjective ratings, and of 84.7 %, based on the user ratings of the social media website, are achieved.**

I. Introduction

The automatic assessment of singing quality is of great relevance in the field of Music Information Retrieval (MIR). Until recently, rating of singing has been the domain of examiners at conservatories or classical singing competitions, but more and more talent shows like ‘Star Search’, ‘Popstars’ or ‘The Voice of Germany’ are flooding the TV market where the audience is often involved in the appraisal of the singers. This trend and the availability of social media websites, such as *YouTube*[1], has resulted in a growing number of people recording and publishing their singing. For their audience, the increasing amount of content makes a manual mining of all items of potential interest impossible and calls for an automated preselection of high-quality recordings. The main application of the technology investigated in this contribution is the automatic rating of singing recordings which are uploaded to social media websites. Furthermore, the proposed system is able to provide feedback on the overall quality to amateurs.

Non-professionals usually rate singing on instinct without knowing the exact reasons. In the entertainment sector, karaoke systems like ‘SingStar’[2] exist which give a rating of each singer based on how exact the given pitch and rhythm are met. However, aesthetical aspects, which are the gist of a musical performance, are not taken into account in these games.

The human voice is considered as the oldest musical instrument and its functional range exceeds those of other instruments. The resonance filters of the vocal tract, i. e., the formants, can be shaped to a large extent, depending on the skill of the singer, and numerous timbres and vocal registers, such as, e. g., *falsetto* or *voce faringea* [1], can be used. As it is the case in appraisal of all kinds of art, the aesthetical perception of singing is very subjective and varies between different referees, so that even experts often disagree on the perfection of a certain performance. This makes an automatic classification of singing quality difficult as we require labels in order to apply supervised machine learning schemes. In our case, we do not have a distinct *ground truth* of labels, i. e., no objectively correct ratings, but a *gold standard* derived from subjective ratings by several referees, or annotators.

In this contribution, we want to present methods for automatic assessment of perceived quality of singing performances on the basis of non-professional, private home-recordings of Western classical and pop music. The goal is to build a system which works for different genres, such as the two mentioned. We restrict ourselves to unisonous, unaccompanied vocal recordings from YouTube. Acoustic features are extracted from the signal on segment-level and then fed into different classifiers. Except for chunking, no signal enhancement or preprocessing is applied.

The contribution is organised as follows: In the next section, we reference related works in the field of singing voice quality and in section III, we gather the music-theoretical background of vocals with particular attention on parameters which might be relevant for perceived quality. In section IV, we describe the used singing corpus and how we collected annotations by professionals and non-professionals to get subjective ratings. Section V presents how those parameters are extracted from the audio signal and which of them have proven to be

[1] https://www.youtube.com/

[2] https://www.singstar.com/home.html

meaningful for the task at hand. The results of classification are shown in section VI before we conclude in section VII.

II. Related Works

Already in 1925, Seashore and Metfessel developed acoustic parameters for quantitative measurement of singing quality [2]. They argue that the main difficulty in finding those parameters is that the crucial point of what makes singing art is not the exact adherence of rhythm and pitches, but the small deviations from the score. Some meaningful characteristics of 'good voice quality', e. g., a stable *vibrato*, have been defined by Bartholomew [3].

Wapnik and Ekholm present a study on inter- and intra-rater reliability in vocal performance evaluation based on 12 criteria that are well-established for the judgement of classical singing [4]. These criteria include *vibrato*, *timbre*, *intonation*, *dynamic range*, *breath control*, and *intensity*. They found out that inter-rater reliability is highest for the *overall score* and the accuracy of *intonation* and that one unfulfilled criterion is correlated with a decrease in the rating w. r. t. all other criteria.

In the meantime, much research has been carried out on certain of the mentioned aspects rather than on a general evaluation of singing. Concerning *vibrato*, Hirano and Sundberg introduce four parameters which are relevant for its quality [5]: *frequency*, *modulation depth*, *regularity*, and *waveform*. However, the recognition of vibrato depends highly on the accuracy of pitch detection, which is quite challenging in case of polyphonic music [6]. Weninger et al. propose a robust approach to automatic recognition of vibrato in polyphonic music [7]. Instead of the discrete fundamental frequency (F0) spectrum, they take into account percentiles of delta regression coefficients of the F0 contour. With this improved vibrato recognition method, the accuracy was higher than 85 % for the examined real-life database.

For evaluation of *singing voice quality*, the *singing power ratio (SPR)* can be used [8]. SPR gives the ratio of the maximum energy in the frequency range of the *singing formant* (2-4 kHz, also called 'singer's formant') to the maximum energy in the range of 0-2 kHz. Watts et al. also investigate if this parameter can be used to estimate the talent of a singer. They discovered that SPR significantly differs between untrained talented and untrained untalented singers.

Concerning untrained singers, criteria for rating can differ from those for trained singers. Cao et al. examine recordings of untrained singers with respect to *intonation*, *vibrato*, *rhythm*, *timbre*, *dynamics*, and *clarity of voice* and the influence of these features on the ratings from experts, who had to judge every criterion [9]. *Intonation* prove to have the highest impact on the ratings, whereas *vibrato* had almost no influence. This might be due to the lack of vibrato in recordings of unprofessional singers. *Rhythm*, *timbre*, and *clarity of voice* had an impact on the rating.

For evaluation of *intonation*, there are basically two different approaches:

1) The estimated F0 contour is compared to the frequencies which correspond to the expected notes [10], [11]. This approach requires the score of the musical piece in digital format, e. g., MIDI. The estimated F0 contour is then simply compared to the prescribed pitch contour using the frequency-independent measure of *cent* (see section III).
2) If the transcription of the musical piece is unknown, intonation must be rated independent from melody. Nakano et al. propose a method to evaluate intonation based on the pitch interval accuracy [12]. The overall stability of intonation is assessed by measuring the pitch offset from an equal tempered scale with a fixed reference tone. This is based on the assumption that this offset is stable for good singers.

Further approaches exist as the one proposed by Mauch et al. [13], [14]. While the *interval error* can be measured in a straightforward way, the computation of *pitch error* requires the knowledge of the reference tone, which is usually not stable in unaccompanied singing. In the proposed method, a 'normalised' representation of the pitch contour is obtained first by removing the nominal pitch in the score. Then, the tuning reference is estimated via either linear regression or a sliding window. Moreover, the authors state that *note duration* has a notable impact on *pitch accuracy*. A good overview of features for tonal analysis of music, e. g., the recognition of played notes and chords, is provided in [15].

Besides the tonal characteristics of the voice, also the analysis of rhythm is of importance for the evaluation of singing, and more generally, music. For the assessment of rhythm, numerous methods exist based on recognition of *beat*, *onsets*, and *tempo*. Onset detection is usually the basis for all other rhythm-related features [16], [17], [18]. However, also an analysis based on the periodicity of the envelope of the audio signal has shown evidence of its practicability [19].

For the assessment of singing skill, several of the mentioned approaches are combined. For example, Nakano et al. fuse features derived from *intonation* and *vibrato* to classify recordings of professional singers as either 'good' or 'bad' [12] and achieve an accuracy of up to 87 %, depending on the gender of the singer.

Many real-time applications already exist for this kind of research, e. g., Gkiokas et al. [20] describe a visual feedback system to assess the *quality of tone* of clarinets and point out the deficits. *InTune* [21] visualises the deviations of pitch from a given reference in singing.

Mayor et al. developed a method for Karaoke systems to analyse the performance of a singer based on a reference score giving also feedback about musical expression [22]. They segment the notes at an intra-note level using *hidden Markov models*. Another study showing that automatic rating of Karaoke singing is close to human rating has been published by Tsai and Lee [23]. Music performance games exist, such as, e. g., *Songs2See* [24], where the musician gets feedback on the accuracy of the played notes. Moreover, Han et al. present a system for musical performance evaluation based on *intonation* [25].

There are, however, only few systems that do not require the musical score as foundation for assessment, such as the system developed by Nakano et al. One approach has been presented by Nichols et al. [26] where they present a method of ranking large amounts of 'home singing' videos for searching talented musicians in YouTube videos. They propose *intonation histograms* and use the most frequent pitch as a tuning reference from which they induce an equal-tempered scale. Besides the deviations from this performance-specific scale, melody-based metrics are extracted as features. An accuracy of up to 67.5 % for pairwise ranking of singing quality is achieved. However, if only the proposed intonation-based features are used, the accuracy is 51.9 %.

In this contribution, we propose new features related to *intonation* for the case of an unknown score.

III. Music-theoretical Background

Here, we focus on Western classical and pop music. This is important to note, because music from other cultures often uses other scales, e. g., oriental music, and in other musical genres, particular singing techniques or vocalisation styles are common, e. g., *death growl* in heavy metal. In general, singing is a discipline of art, so both interpretation and aesthetical perception are subjective. Thus, there is no distinct agreement on which criteria determine the quality of singing.

In the following, all criteria that are taken into account in our work, are described.

A. Intonation

There is much evidence that a major criterion of good vocals is intonation [4], [9]. *Clean* intonation means that the singer hits and holds out the correct notes, i. e., the pitch matches exactly the one the singer intends to sing. However, intonation is also an essential mean of *musical expression*. As an example, a tone is often sung too sharp, i. e., its pitch is a little too high, before dissolving the arc of suspense.

In principle, to decide whether one note is sung in a clean way, a reference is required. There might be three reasons why this reference is not available:

1) There is no general reference for pitch in musical scales. The standard pitch of the note $A4$ has been standardised to 440 Hz in ISO 16:1975; however, orchestras usually do not stick to this agreement and choose standard pitches between 435 Hz and 445 Hz. If singers are unaccompanied, they usually choose a reference which deviates much more from standard pitch as most humans have no *absolute pitch*.
2) There are several intonation systems. The intonation system defines the frequency ratios between the 12 notes (which is standard in Western music) within one octave. In *pure intonation*, those frequency ratios are of small whole numbers, i. e., the harmonic series [27]. This is why the intervals in pure intonation are perceived as clean. However, if several pure intervals are combined successively, the resulting interval between the first and the last note is usually not pure. For singers, adaptation of pitches to sing pure intervals preferentially is common practice, whereas this is not the case for instruments with fixed pitches, like piano. For this reason, the *equal temperament* has been established as a standard tuning in Western music, which divides one octave into 12 intervals of relatively equal width (semitones) [28]. This means that, with the interval measure of *cent* between two frequencies f_1 and f_2, $i = \log_2 \frac{f_1}{f_2} \cdot 1200\ \mathrm{cent}$, all semitones have an equal distance of 100 cent.
3) In addition to that, our goal is to build a system which does not require a transcription of the notes of the musical piece. Thus, the melody, i. e., the sequence of intervals the singer pursues to sing, is also unknown.

B. Voice Quality

Besides intonation, a fundamental question is what discriminates the voice of a 'good' singer against the voice of a 'bad' singer. Two important criteria are *vibrato* and *timbre*.

Vibrato is a continuous oscillation of the pitch of a tone which is held out. Frequencies (rates) of vibrato around 6 Hz are perceived as pleasant by human listeners [29].

One major aspect of *timbre* is the *singing formant*, which is located in the band around 3 000 Hz and is of importance for the assertiveness of the voice in an orchestra [30]. This is due to the high sensitivity of the human ear at these frequencies. Furthermore, the *clarity of voice* is an important criterion. Untrained singers often have a 'breathing' or 'aspirating' voice, because their vocal cords do not close properly. Such deficiencies in breath control lead to a dull voice, caused by non-harmonic signal parts. However, an overemphasis of harmonics leads to a sharp sound, which is not desirable either. So, a well-balanced ratio of harmonic and non-harmonic parts is essential for a good sound and a lively voice [31].

C. Dynamics

Dynamics, i. e., variations of loudness, are another criterion for *musical expression* and the arrangement of a musical performance. A constant loudness throughout a musical piece, especially in classical music, can be boring. With meaningful dynamics, the artist can evoke emotions in the audience.

IV. Singing Corpus

Let us now turn to our study. For our experiments, we created a corpus out of private non-professional home recordings from the video-sharing website YouTube.

We selected performances of seven popular songs given in table I, from which several recordings in different degrees of quality are available. Only one singer is present in each video and the singing is unaccompanied; however, different kinds of noises and disturbances are found. All exploited recordings have been published under a *Creative Commons CC BY Licence*[2]. Table I shows also the number of samples of each song and the gender distribution.

[2] https://creativecommons.org/licenses/by/3.0/

Song	Author	male	female	overall
Amazing Grace	William Walker	4	8	12
Ave Maria	Franz Schubert	8	15	23
Someone like you	Adele	2	8	10
Hallelujah	Leonard Cohen	5	8	13
Over the Rainbow	Harold Arlen	2	22	24
The Star-Spangled Banner	John Stafford Smith	8	17	25
Time to Say Goodbye	Francesco Sartori	2	1	3
overall		**31**	**79**	**110**

TABLE I
LIST OF THE SONGS, NUMBER OF SAMPLES, AND DISTRIBUTION OF GENDER FOR EACH SONG

The audio tracks were segmented without prior processing. The beginning of each recording was cropped until the entry of voice and the succeeding minute was divided into two segments of 30 seconds each, called *snippets* in the following. As there is one video clip with only 35 seconds of singing, we gained a corpus of 219 snippets out of 110 recordings.

A. Ratings

In order to apply supervised learning schemes for audio classification, we need class labels, or ratings, for each snippet. As *perceived singing quality* is a subjective matter and there is no objectively 'correct' classification, it is necessary to have a multitude of annotators [32]. We employed three different methods to obtain ratings:

1) We collected subjective ratings of the audio by means of a web application called *Record Ratings* based on the web framework *Ruby on Rails*[3]. This application plays back the audio-only snippets randomly and asks the annotator to give a grade as a number of *stars* between 1 and 10. The annotators were given the three following questions to form their opinion: *"How much do you like the singing voice?"*, *"Is the singing out of tune or not?"*, and *"Does the singing transport emotion or is it horribly boring?"*. It was pointed out that the annotators should not rate neither the technical quality of the recording nor the song itself.
   For statistical reasons, each annotator had to give information on his or her age, gender, if he/she plays an instrument, sings, has had musical education or is a professional musician. In the first series, 6 annotators were asked to rate all snippets of the corpus with the given application. The group consisted of 3 female and 3 male subjects aged 24 to 59 with different musical background (see table II). There was one complete layperson, all others have had musical education, play an instrument, and two of them also sing professionally.

   Inter-rater reliability in terms of *Krippendorf's Alpha*, which takes the order of discrete scales into account, is 0.313. This means that the consensus on the quality of the singing is quite low and maintains the assumption that assessment of singing quality is a highly subjective task. Our data cannot approve that this is due to a generally worse rating by the professional musicians. E. g., annotator D is a professional musician and singer whereas annotator E is a layperson and the distributions of their ratings are quite similar. The only correlation we found is that older annotators tend to give worse grades. Deviations in the judgement can also depend on the playback device. However, in 73 % of the snippets, the difference in the rating of both snippets from one singer differs not or only in one *star* (out of ten).

Annotator	Gender	Age	Instr.	Singer	Education	Professional
A	female	59	yes	no	yes	yes
B	female	24	yes	yes	yes	yes
C	female	27	yes	no	yes	no
D	male	24	yes	yes	yes	yes
E	male	27	no	no	no	no
F	male	59	yes	no	yes	yes

TABLE II
INFORMATION (GENDER, AGE, INSTRUMENTALIST, TRAINED SINGER, MUSICAL EDUCATION, PROFESSIONAL MUSICIAN) ON THE SIX ANNOTATORS OF THE FIRST SERIES OF ANNOTATIONS

2) In the second series, *Record Ratings* was used as a *crowdsourcing* platform. Anybody was allowed to sign up and was given 25 randomly chosen snippets to rate. In total, 96 annotators participated (55 female, 41 male) aged 21 to 77. The distribution of expert knowledge was relatively equal, but only 16 of them were complete laymen. It is also possible that the same user participated several times in the experiment or skipped some of the 25 snippets. In total, 2 197 ratings of snippets were collected by crowdsourcing.
3) On YouTube, the number of *views*, *likes*, and *dislikes* is available and it is feasible to transfer those three numbers into one metric or discrete classes of quality. However, the drawback of this labelling method is that ratings might refer to the quality of the recording itself or the visual nature of the singer rather than the singing. The numbers of views, likes, and dislikes for each song are summarised in table III.
   As a target label based on these measures, the following two metrics have been evaluated:

$$\text{YouTube1} = \max\left(0, \log_2\left(\frac{\#\text{likes}}{\#\text{dislikes}} \cdot \#\text{views}\right)\right)$$
$$\text{YouTube2} = \max\left(0, \log_2\left(\frac{\#\text{views}}{\#\text{likes}}\right)\right)$$

   The logarithm was taken to have a denser region of values, which usually leads to a better classification performance [32].

The resulting quality measures of each approach were mapped to three discrete classes of quality: **poor**, **fair**, and **good**.

For the ratings obtained by *Record Ratings*, grades 1 to 3 were mapped to quality *poor*, grades 4 to 6 were mapped to

[3] http://rubyonrails.org/

Song	Views			Likes			Dislikes		
	min.	max.	avg.	min.	max.	avg.	min.	max.	avg.
Amazing Grace	78	698 989	115 111	0	2 146	521	0	48	12
Ave Maria	17	40 456	5 347	0	633	44	0	11	1
Hallelujah	29	1 796	521	0	23	6	0	4	1
Over the Rainbow	19	9 793	761	0	68	9	0	8	1
Someone like you	92	8 959	1 749	0	277	35	0	9	2
The Star-Spangled Banner	15	352 791	15 321	0	665	33	0	43	3
Time to Say Goodbye	54	1 099	550	2	12	7	0	2	1
overall	15	698 989	21 923	0	2 146	79	0	48	3

TABLE III
NUMBER OF MINIMUM, MAXIMUM, AND AVERAGE VIEWS, LIKES, AND DISLIKES PER SONG

Class	Ranges of YouTube1	Ranges of YouTube2
poor	(0,8.0]	(8.0,∞)
fair	(8.0,16.0]	(5.5,8.0]
good	(16.0,∞)	[0,5.5]

TABLE IV
MAPPINGS BETWEEN RANGES OF METRICS YOUTUBE1 & YOUTUBE2, AND THE THREE CLASSES

Class	Majority vote	YouTube1	YouTube2
poor	89	76	40
fair	95	108	118
good	35	35	61

TABLE V
DISTRIBUTION OF CLASSES

quality *fair*, and grades 7 to 10 were mapped to quality *good*. The last class covers more grades as they are much less present in the ratings than the lower grades. A *gold standard*, i. e., the labels used as targets for supervised learning, was then defined as the majority of the classes present in the combined ratings of both series (see 1) & 2)) with *Record Ratings* (overall 3 051 ratings).

The metrics based on YouTube statistics were mapped to classes intuitively according to table IV. Note that it is considered that a smaller value in the measure YouTube2 implies better singing quality.

Finally, we end up with nine different target labels for each snippet: Six ratings from the single annotators, one originating from the majority vote of these annotators and crowdsourcing and two ratings originating from YouTube statistics.

The class distribution for majority vote and the two metrics from YouTube are shown in table V. For all target labels, the most frequent class is *fair* and the least frequent quality is *good*.

V. METHODOLOGY

In this section, we describe first the acoustic features which are extracted from the audio signals and used for classification. This includes both standard features and intonation-based features which have been implemented for the research resulting in this contribution. We then point out the results of feature selection and specify the employed classifiers.

A. Standard Feature Sets

The extraction of acoustic standard features is done by the tool OPENSMILE [33]. OPENSMILE provides, among other things, the computation of low-level descriptor (LLD) contours such as loudness, pitch, MFCC, jitter & shimmer, and the computation of functionals of these LLDs such as mean, moments, percentiles, etc.

In our experiments, the baseline feature set from the *INTERSPEECH 2013 Computational Paralinguistics Challenge (ComParE)* [34] was used. It comprises functionals of 60 LLD contours and their 1st & 2nd order derivatives, in total 6 373 acoustic features per audio segment or snippet. The whole lists of features are given in [34]; we now limit ourselves on the description of those features in the set which are most relevant to human voice, as pointed out in section III.

The *fundamental frequency (F0)* or *pitch* of the voice signal is mainly responsible for the perceived note. Thus, the extraction of pitch contour is also the basis for the later introduced intonation-based features. The computation is done using the *Subharmonic Summation (SHS)* method, where F0 can also be detected when only its harmonics are present in the signal, and afterwards *Viterbi smoothing* is applied. It is important to note that the search range for F0 must be adjusted to 16 Hz – 1 400 Hz, which includes the human pitch range in singing, as the standard configuration covers only pitches typical for speech.

Harmonics-to-Noise-Ratio (HNR) gives information on the amount of noise in a periodic audio signal. As stated in section III; it is an indicator for *voice quality*. For HNR, the *autocorrelation function (ACF)* is computed and its first peak, which is at the fundamental period, is set in relation to the overall signal energy [32].

MFCCs are a well-established acoustic feature, which takes the subjective scale of human hearing into account and is capable of separating the excitation part from the resonance filter part in the speech signal [32]. In COMPARE, coefficients 1 to 14 are used.

The *auditory spectrum* models the human loudness perception of different frequencies and also takes the summation within critical bands into account [35]. Features based on the auditory spectrum are therefore able to reflect the human perception of, e. g., *timbre*, in a better way.

B. Intonation-based Features

In addition to the described well-established acoustic features, features based on intonation have been implemented prototypically in *Matlab*.
As a first step, the F0 contour is extracted with *openSMILE*, using a Gaussian window with a width of 100 ms and a hop size of 10 ms. To evaluate how well a tone is hit, a reference must be determined. Multiple references are tried out:

1) Standard pitch A4 (440 Hz).
2) A set of ten *candidates* extracted from the audio sample itself. The deviation is computed for each one of the candidate references, but only the candidate with the lowest deviation is included in the final feature vector.
The determination of the *candidates* is similar to the method proposed in [26]. A *tolerance* is defined beforehand to define the maximum deviation of pitch to be interpreted as the same tone. A tolerance of ±20 cents was chosen as the deviation of, e. g., a *minor third* between *pure intonation* and *equal temperament* is 16 cents. So, the almost inaudible difference of the tuning system does not affect the proposed features. A histogram of the extracted pitch (F0) is created, where each pitch value is assigned to one 'interval of tolerance'. The centre frequencies of these intervals are created from the F0 contour itself. Each time an F0 occurs which does not fit into any existing interval, a new interval is added to a list. Finally, the mean of the F0 values of the ten most frequent intervals are chosen as candidates for reference.

Using standard pitch A4 and each of the ten candidates as a reference f_{ref}, *semitone scales* are computed according to the formula

$$f(i) = f_{\mathrm{ref}} \cdot 2^{\frac{i}{12}}, \tag{1}$$

where i takes all integers so that the resulting frequency is in the range of human pitch (see section V-A).

Now, it would not be wise to compute the error in intonation for every single frame, as *vibrato*, which is desired, causes changes in pitch. So, the next step is *segmentation*, where several frames belonging to one note are combined. As a first step, the semitone (based on the absolute scale derived from standard pitch and the scales derived from all candidates) closest to F0 is determined for each frame. If the assigned semitone is different from that of the preceding F0, or if the pitch detection failed, a new segment is started, otherwise, F0 is added to the current segment. Without any post-processing, short segments consisting of only one or two frames would be generated. So, as a second step of segmentation, to improve robustness, a minimum segment length of 3 frames is introduced. Shorter segments are combined with the adjacent segment that has smaller distance in cents. Fig. 1 exemplifies the segmentation of 13 frames into 3 segments. This results in a sequence of the sung tones, thus, this intermediate result is a very basic transcription of the melody.

Now, to determine *intonation*, three different approaches were pursued.

1) For the first set of intonation-based features, the error of intonation is computed frame-wise, i. e., the segmentation is not considered here. For each frame and each reference semitone scale (equation 1), the minimum distance of the current pitch and all pitches from the respective semitone scale in cents are used as *intonation error*. Functionals according to table VI are applied to obtain a feature vector for the whole snippet.
2) To obtain features, which are more robust against intended deviations in pitch, such as vibrato, the results of segmentation are used. The error between the pitch of each frame within a segment and the semitone assigned to the segment is computed (in cents). To obtain measures on segment-level, five functionals shown in table VI are applied on the sequence of errors within each segment.
3) As a third approach, we now look at the musical intervals between the segments. Assuming that a singer cannot hold the same reference over the whole song, e. g., the pitch gets higher and higher, this would not affect the subjective perception of intonation, usually, if the rise is not too fast. However, the reference tone of our error measures does not change throughout the piece, so this would have an impact on the computed features.
The musical interval is, in principle, the frequency ratio between two adjacent segments. The derived features are based on the mean pitch of each segment. We look at nine different orders, i. e., the deviations between one segment and its nine successors are computed. The differences of the mean pitches of two segments are now taken, and a modulo 100 operation is executed as deviations of multiples of 100 cents (semitones) are clean intonations, as well as deviations within a tolerance range. Thus, errors of less than ±20 cents are set to 0. The functionals shown in table VI are applied to all deviations of the same order.

As mentioned, both the standard pitch A4 and the 10 candidates are used as a reference f_{ref} in all three approaches, but for the candidates, only the sequence of errors with the lowest mean error is kept, as the alternative 9 candidates do not appear to be appropriate.

Overall, we end up in 112 intonation-based features listed in table VI.

C. Feature Selection

The feature space of COMPARE is quite large, which usually results in reduced classification performance. To reduce its dimension, we employ entropy-based features selection techniques [36], in particular *information gain* and *gain ratio*.

Feature selection was performed on the whole corpus introduced in section IV, where the majority vote of the listeners

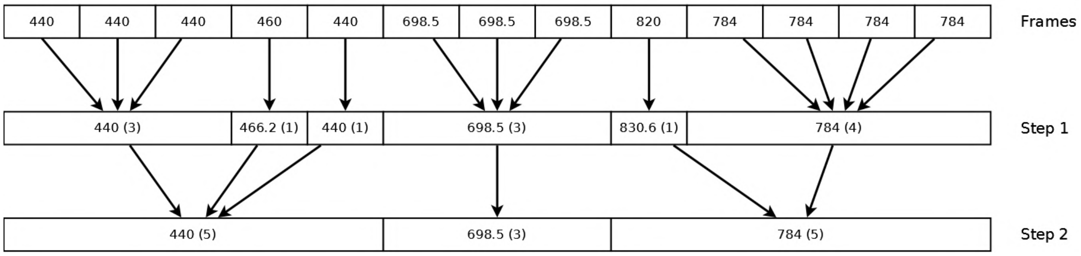

Fig. 1. Example of the segmentation procedure. Given numbers are pitches in Hz. Numbers in brackets are sizes of the segments (number of frames).

Dim.	LLD	Functionals
10	Frame-wise deviation from semitone scale (Reference pitch: $A4$ & candidate with minimum error)	Mean, maximum standard deviation, mean squared error, percentage of frames with deviation above tolerance
30	Deviation from semitone scale per segment (Reference pitch: $A4$ & candidate with minimum error)	*Segment level:* Mean, maximum, standard deviation, mean squared error, ratio of maximum pitch and the sum of the other pitches *Snippet level:* Mean, standard deviation, maximum error
72	Deviation of intervals between segments (1^{st} to 9^{th} order) (Reference pitch: $A4$ & candidate with minimum error)	Mean, maximum standard deviation, percentage of frames with deviation above tolerance

TABLE VI
OVERVIEW OF THE INTONATION-BASED LLDS AND THEIR FUNCTIONALS

is used as target label. It was found that MFCC, which are supposed to correlate with *timbre*, have a high rank, as well as loudness. The other voice quality feature, HNR, is not very relevant in the results of feature analysis.

Three subsets of ComParE were defined:

- ComParE_50, consisting of the 50 features with highest rank according to *information gain*
- ComParE_30, consisting of the 30 features with highest rank according to *gain ratio*
- ComParE_reduced (ComParE_red.), consisting of 50 manually selected features (inspired by the music-theoretical background, including HNR and harmonicity).

D. Classifiers

In our experiments, we evaluated the performance of four different classifiers, in order to ensure that our findings w. r. t. the optimum feature set are independent from the employed machine learning scheme.

- Support vector machines (SVM), implemented with *Sequential Minimal Optimisation (SMO)* [37], with *linear* and *radial basis function (RBF)* kernel, and the *complexity parameter C*,
- Naive Bayes (NB) [38],
- K-Nearest Neighbours (kNN), with *linear search*, *Euclidean distance* and different numbers of neighbours k [39],
- A C4.5 decision tree (C4.5) [40] with different confidence values C.

All classifiers were trained and evaluated using the data mining software *Weka*[4] [41].

VI. Evaluation

As our corpus (see section IV) is not large enough to be split up into a training and a test partition, we employed two modified versions of leave-one-out cross validation:
In *leave-one-singer-out cross validation* (LOSiO-CV), the model is iteratively trained on all snippets except for the two snippets of one singer, on which the model is then evaluated.
In *leave-one-song-out cross validation* (LOSoO-CV), the model is iteratively trained on all snippets of the 6 training songs and then evaluated on the snippets of the remaining song.
This procedure promises results which are more realistic with regard to singers or songs that are not included in our corpus. As the three quality classes are not *balanced*, training instances of less represented classes are upsampled before each model training. As quality measure, *unweighted average recall (UAR)* is used. The UAR is defined as the average *recall* over all classes, where the class-specific recall is given as the ratio of the number of correctly classified snippets from the respective class and the total number of snippets in this class. This is ideal in case of unbalanced instance per class distribution. The chance level of the UAR in a 3-class learning task, i. e., the performance achieved by a classifier with random or constant predictions, is $\frac{1}{3} \approx 33.3\,\%$.

In the following, the results of the evaluation are shown separately for each target introduced in section IV: the six single annotators, majority vote (of the ratings from six annotators and crowdsourcing), and the two measures from YouTube ratings. For each case, the result with the CV method leading to the superior result is shown.

A. Single Annotators

An evaluation of rater-dependent models is justified by the relatively low inter-rater reliability (see section IV). Classi-

[4] http://www.cs.waikato.ac.nz/ml/weka/

Annotator	Feature Set	Classifier	C	UAR
A	ComParE_30	SVM (linear)	0.06	69.8 %
	Intonation	SVM (RBF)	0.8	50.8 %
	ComParE_red. + Intonation	SVM (RBF)	0.05	52.8 %*
B	ComParE_red.	SVM (linear)	1.2	51.1 %
	Intonation	SVM (linear)	1.6	43.6 %
	ComParE_red. + Intonation	SVM (linear)	1.2	49.5 %
C	ComParE_50	SVM (linear)	1.8	48.6 %
	Intonation	SVM (linear)	1.0	45.9 %
	ComParE_50 + Intonation	SVM (linear)	0.2	44.5 %
D	ComParE_50	SVM (linear)	0.1	55.4 %
	Intonation	SVM (linear)	0.4	44.3 %
	ComParE_50 + Intonation	SVM (linear)	1.4	53.2 %
E	ComParE_red.	SVM (linear)	0.1	44.6 %
	Intonation	SVM (RBF)	1.4	42.6 %*
	ComParE_red. + Intonation	SVM (linear)	0.1	44.9 %
F	ComParE_red.	SVM (linear)	0.7	62.5 %
	Intonation	SVM (linear)	0.04	52.1 %
	ComParE_red. + Intonation	SVM (linear)	0.07	57.5 %

TABLE VII
Best classifiers for each annotator and each type of feature set. *Evaluated with LOSoO-CV, all other classifiers were evaluated with LOSiO-CV

fiers were trained for each annotator (A-F), and for each ComParE-based feature set, the intonation-based feature set (Intonation) and combinations of the ComParE-based feature sets and intonation-based features. Table VII shows the results in terms of UAR for all annotators and each of the three categories of feature sets. The best performance is achieved with SVM in all cases; the UAR of 69.8 %, with annotator A and ComParE_30, is maximum. The best results with the other classifiers in terms of UAR are: 60.2 % with kNN, 57.7 % with NB and 47.3 % with a C4.5 decision tree.

It can be observed that the classification performance for annotators C and E are worse than for the other annotators. These two annotators are those who are not professional musicians. This can be a clue that the annotations by experts are more consistent than those of laymen, but we cannot draw a final conclusion on that based on only two non-professionals. Furthermore, as further investigation of the raters' decisions shows, the 10-grade scale has been exploited differently by all annotators. A normalisation of the ratings might have improved the consistency of the classification results between different annotators.

For the ComParE-based features, the reduced sets always led to better results than the whole set. The intonation-based features alone have an average performance of only 46.5 % UAR and also the combined features are usually worse than the pure ComParE features. However, the ratings are not only based on intonation, so we do not have meaningful targets to get a final conclusion on the relevance of the proposed intonation-based features. A feature selection on these features has not been performed and might improve the performance. However, we decided to present only results with the full intonation-based feature set. The main reason for that is that feature selection on large redundant acoustic feature sets usually results in a rather arbitrary subset. Moreover, feature reduction and classifier would be trained on the same data set. Thus, we would not be able to draw a meaningful conclusion.

Classifier	Parameter	Feature Set	UAR
SVM (linear)	$C = 0.9$	ComParE_30	55.0 %
	$C = 0.3$	Intonation	37.6 %*
	$C = 0.2$	ComParE_red. + Intonation	55.4 %
kNN	$k = 14$	ComParE_red.	59.3 %
	$k = 20$	Intonation	35.6 %*
	$k = 20$	ComParE_50 + Intonation	47.8 %*
NB		ComParE_50	52.9 %
		Intonation	29.9 %
		ComParE_50 + Intonation	49.5 %*
C4.5	$C = 0.5$	ComParE_red.	47.6 %*
	$C = 0.5$	Intonation	37.0 %
	$C = 0.4$	ComParE_30 + Intonation	43.5 %

TABLE VIII
Best results per classifier and each type of feature set for target labels from majority vote. *Evaluated with LOSoO-CV, all other classifiers were evaluated with LOSiO-CV

B. Majority Vote

Table VIII shows the results for all four evaluated machine learning schemes and all three types of feature sets for the target labels generated from the majority vote of all annotators. The highest recognition rate of 59.3 % UAR is achieved with kNN. This differs from our result for single annotators, where SVM works best in all cases. The UAR with SVM is only 55.4 % here. The results with naive Bayes and a Decision Tree are 52.9 % and 47.6 % UAR, respectively. The poor performance for naive Bayes might be due to the fact that the single features are not statistically independent and naive Bayes is not able to cope with redundancies very well. Decision trees are generally very powerful if there is a small subset of meaningful features, which is probably not the case for this task.

Concerning the feature sets, the best result is attained with ComParE_reduced. However, with SVM, the result of combined intonation-based and reduced ComParE features is slightly better, whereas this is not at all the case for kNN. Intonation-based features only yield the worst results with all classifiers.

A UAR of 59.3 % is certainly much better than 'random classification' (33.3 % UAR), but as well worse than desired. The crucial point is that we depend completely on the quality of annotations. If the majority vote is done iteratively without the ratings of one annotator, and the majority is then compared to the remaining annotator, the UAR is only 52.7 % on average. This means that the inter-rater reliability is not very high. Thus, a UAR of 59.3 % is still considerable.

C. YouTube Ratings

Table IX shows the best results for both target classes generated from the ratings on YouTube. For the targets which take also *dislikes* into account (YouTube 1), the maximum UAR is 49.13 %, for the ratio of *likes* and *views* only (YouTube 2), the maximum is 84.7 %, achieved with SVM and a combination of intonation-based features and ComParE_30. This is by far the best performance in all our experiments.

Target	Feature Set	Classifier	Parameter	UAR
YouTube1	ComParE_red.	kNN	$k = 19$	49.1 %
	Intonation	NB		48.6 %
	ComParE_red. + Inton.	NB		45.6 %
YouTube2	ComParE	SVM (lin.)	$C = 10^{-5}$	51.1 %
	Intonation	SVM (lin.)	$C = 0.04$	52.7 %
	ComParE_30 + Intonation	SVM (lin.)	$C = 0.01$	84.7 %

TABLE IX
Best results for all types of feature sets and both target classes generated from YouTube ratings (LOSIO-CV)

Now, the question arises, why the classification performance is much better than with the gold standard created from the majority vote of the annotators. One possible explanation is that the number of ratings is much larger on YouTube than on the *Record Ratings* platform, which results in a higher consistency. It could nevertheless be the case that the number of *likes* is larger if the recording quality is better. As the ComParE_30 feature set comprises also LLDs which are common in general audio classification, it is possible that the general audio quality has also been modelled in the proposed system.

Finally, it must be pointed out that the relatively low number of instances in the corpus and the large number of features can result in over-fitting, i. e., a model adapts to the given data too tightly and would not work with a similar accuracy on new, unseen data or different corpora. The easiest way to tackle this problem is simply to collect more labelled recordings. However, from our point of view, with a feature vector of size 162 (112 intonation based features + 50 (ComParE features), there is no disproportion in consideration of a data set of 219 instances. In the INTERSPEECH ComParE tasks [42], [34] and also in the MediaEval challenge [43], it has been shown that a large feature vector of more than 6 000 features led to an improvement for several speech-based recognition tasks compared to a reduced number of acoustic features. This applies even though the features are highly redundant.

VII. Conclusions and Outlook

The best classification performance in terms of UAR was achieved on ratings generated from the ratio of *likes* and *views* on YouTube. For each of the three classes *poor*, *fair*, and *good*, 84.7 % of the snippets were classified correctly on average. The accuracy based on targets from majority vote of annotators is only 59.3 % UAR. This might be due to a too small number of participants (six) for annotation in combination with the high task subjectivity.

Concerning features for the assessment of singing quality, MFCCs, and loudness have proven to be quite meaningful, but there is no final conclusion on a specific feature set. Overall, SVM seems to be the most appropriate machine learning scheme among the considered ones, besides kNN.

While the intonation-based features have not been very beneficial in classification based on manual annotations of singing quality, the best result on YouTube ratings has been achieved with a combination of the 30 ComParE features with highest rank from *gain ratio* analysis and 112 intonation-based features. Those intonation-based features alone yield a UAR of up to 52.7 %. This exceeds the results presented in [26] of 51.9 % in a 2-class decision, although it is certainly delicate to compare results of two distinct approaches on different datasets. Also the performance using an augmented feature set in our 3-class problem (59.3 %) seems better than that of the system proposed by Nichols et al. (67.8 %) for 2-class decisions.

Overall, the results show that the largest room for improvement is now in reaching a more reliable gold standard by a larger amount of ratings. We have found evidence that a model based on a larger number of annotators usually works more robust than a model based on the annotations from only a few experts. This finding is common for many subjective machine learning tasks, such as, e. g., affect recognition or speaker likeability. Furthermore, it would be interesting to let experts rate on different criteria separately, e. g., dynamics, musical expression, and intonation. Having a gold standard on intonation would also be helpful to decide which of the proposed intonation-based features are most meaningful.

In future work, other features related to the singing voice need to be evaluated, such as features describing *vibrato* as presented in [7], *rhythm* [18], *singer traits* [44], or *emotion* [45].

Further, recent *deep learning* approaches might be capable of improving the classification accuracy [46], [47]. Besides, other feature representations, such as *bag-of-audio-words* [48] are worthwhile to be investigated in this context. *Active learning* [49] or *cooperative learning* [50] could help to reduce the effort of annotation, considering the loads of YouTube videos that still conceal many potential talents.

Anyway, rating of vocals is still a very subjective task. Although, with the proposed system, it is possible to get a rough automatic assessment of vocals, the proposed system cannot substitute a singing teacher as our system does not tell how to improve the singing technique.

Acknowledgment

The research leading to these results has received funding from the European Union's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu).

References

[1] A. Mayr, "Investigating the voce faringea: Physiological and acoustic characteristics of the bel canto tenor's forgotten singing practice," *Journal of Voice*, 2016, available online.

[2] C. E. Seashore and M. Metfessel, "Deviation from the regular as an art principle," *Proc. National Academy of Sciences USA*, vol. 11, pp. 538–542, 1925.

[3] W. T. Bartholomew, "A physical definition of "good voice quality" in the male voice," *Journal of the Acoustic Society of America*, vol. 6, pp. 25–33, 1934.

[4] J. Wapnik and E. Ekholm, "Expert consensus in solo voice performance evaluation," *Journal of Voice*, vol. 11, no. 4, pp. 429–436, 1997.

[5] M. Hirano and J. Sundberg, *Vibrato*, P. H. Dejonckere, Ed. San Diego: Singular Publishing Group, 1995.

[6] I. Luengo, I. Saratxaga, E. Navas, I. Hernáez, J. Sanchez, and I. Sainz, "Evaluation of pitch detection algorithms under real life conditions," in *Proc. ICASSP*. Honolulu, Hawai: IEEE, 2007, pp. 1057–1060.

[7] F. Weninger, N. Amir, O. Amir, I. Ronen, F. Eyben, and B. Schuller, "Robust feature extraction for automatic recognition of vibrato singing in recorded polyphonic music," in *Proc. ICASSP*. Kyoto, Japan: IEEE, 2012, pp. 85–88.

[8] C. Watts, K. Barnes-Burroughs, J. Estis, and D. Blanton, "The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers," *Journal of Voice*, vol. 20, no. 1, pp. 82–88, 2006.

[9] C. Cao, M. Li, J. Lie, and Y. Yan, "A study on singing performance evaluation criteria for untrained singers," in *Proc. ICSP*, Beijing, China, 2008, pp. 1475–1478.

[10] C. Weihs, U. Ligges, J. Güttner, P. Hasse-Becker, and S. Berghoff, "Classification and clustering of vocal performances," in *Proc. Annual Conference of the Gesellschaft für Klassifikation e.V.*, M. Schader, W. Gaul, and M. Vichi, Eds. Berlin: Springer, 2002, pp. 118–126.

[11] P. Żwan, "Automatic singing quality recognition employing artificial neural networks," *Archives of Acoustics*, vol. Vol. 33, N, pp. 65–71, 2008.

[12] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. INTERSPEECH*. Pittsburgh, PA: ISCA, 2006, pp. 1706–1710.

[13] M. Mauch, K. Frieler, and S. Dixon, "Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory," *Journal of the Acoustic Society of America*, vol. 136, no. 1, pp. 401–411, 2014.

[14] J. Dai, M. Mauch, and S. Dixon, "Analysis of intonation trajectories in solo singing," in *Proc. ISMIR*, Málaga, Spain, 2015, pp. 420–426.

[15] B. Schuller and B. Gollan, "Music theorie and perception-based features for audio key determination," *Journal of New Music Research*, vol. 41, no. 2, pp. 175–193, 2012.

[16] M. Goto, "A real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.

[17] M. Alsonso, B. David, and G. Richard, "Tempo and beat-estimation of musical signals," in *Proc. ISMIR*, Barcelona, Spain, 2004, pp. 158–163.

[18] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer, "On rhythm and general music similarity," in *Proc. ISMIR*, 2009, pp. 525–530.

[19] B. Schuller, F. Eyben, and G. Rigoll, "Tango or waltz?: Putting ballroom dance style into tempo detection," *EURASIP Journal on Audio, Speech and Music Processing*, vol. 2008, no. 6, 2008.

[20] A. Gkiokas, K. Perifanos, and S. Nikolaidis, "Real-time detection and visualization of clarinet bad sounds," in *Proc. DAFx*, Espoo, Finland, 2008, pp. 1–6.

[21] K. A. Lim and C. Raphael, "Intune: A system to support an instrumentalist's visualization of intonation," *Computer Music Journal*, vol. 34, no. 3, pp. 45–55, 2010.

[22] O. Mayor, J. Bonada, and A. Loscos, "Performance analysis and scoring of the singing voice," in *Proc. AES Int. Conf.*, London, UK, 2009, pp. 1–7.

[23] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1233–1243, 2012.

[24] E. Cano, S. Grollmisch, and C. Dittmar, "Songs2See: Towards a new generation of music performance games," in *Proc. CMMR*, London, UK, 2012, pp. 19–22.

[25] Y. Han, S. Kwon, and K. Lee, "A musical performance evaluation system for beginner musician based on real-time score following," in *Proc. International Workshop on Networking Issues in Multimedia Entertainment*, San Jose, CA, 2013, pp. 120–121.

[26] E. Nichols, C. DuHadway, H. Aradhye, and R. F. Lyon, "Automatically discovering talented musicians with acoustic analysis of youtube videos," in *Proc. ICDM*. IEEE, 2012, pp. 559–565.

[27] C. Schmidt-Jones, "Tuning systems," *Connexions (March 2011). http://cnx. org/content/m11639/1.21*, 2005.

[28] M. Kennedy, *The Oxford Dictionary of Music*, 2nd ed., J. Bourne, Ed. Oxford University Print, Oxford, 1994, available online: http://www.oxfordmusiconline.com.

[29] S. Anand, J. M. Wingate, B. Smith, and R. Shrivastav, "Acoustic parameters critical for an appropriate vibrato," *Journal of Voice*, vol. 26, no. 6, pp. 820–e19, 2012.

[30] J. Sundberg, "The acoustics of the singing voice," *Scientific American*, vol. 236, no. 3, pp. 104–116, 1977.

[31] J. Pilaj, *Singen lernen mit dem Computer*. Wißner-Verlag, Augsburg, 2011.

[32] B. Schuller, *Intelligent Audio Analysis*. Springer-Verlag, Berlin, Heidelberg, 2013.

[33] F. Eyben, F. Gross, F. Weninger, and B. Schuller, "Recent developments in openSMILE, the munich open-source mulitmedia feature extractor," in *Proc. ACM MM*, Barcelona, Spain, 2013, pp. 835–838.

[34] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 148–152.

[35] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models*. Springer Science & Business Media, 2007, vol. 22.

[36] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1998.

[37] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1998.

[38] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proc. Conference on Uncertainty in Artificial Intelligence*, Montreal, QU, 1995, pp. 338–345.

[39] D. W. Aha and D. Kibler, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[40] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. Volume 11, no. 1, pp. 10–18, 2009.

[42] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. INTERSPEECH*. Brighton, UK: ISCA, 2009, pp. 312–315.

[43] F. Weninger, F. Eyben, and B. Schuller, "The TUM approach to the MediaEval music emotion task using generic affective audio features," in *Proc. MediaEval*. Barcelona, Spain: CEUR, 2013, 2 pages.

[44] F. Weninger, M. Wöllmer, and B. Schuller, "Automatic assessment of singer traits in popular music: Gender, age, height and race," in *Proc. ISMIR*. Miami, FL: ISMIR, 2011, pp. 37–42.

[45] F. Eyben, G. L. Salomao, J. Sundberg, K. Scherer, and B. Schuller, "Emotion in the singing voice – a deeper look at acoustic features in the light of automatic classification," *EURASIP Journal on Audio, Speech, and Music Processing, Special Issue on Scalable Audio-Content Analysis*, vol. 2015, 2015, 9 pages.

[46] X. Wang and Y. Wang, "Improving content-based and hybrid music recommendation using deep learning," in *Proc. ACM MM*, Orlando, FL, 2014, pp. 627–636.

[47] G. Trigeorgis, F. Ringeval, R. Brückner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*. Shanghai, P. R. China: IEEE, 2016, pp. 5200–5204.

[48] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. INTERSPEECH*. San Francisco, CA: ISCA, 2016, pp. 495–499.

[49] M. Wang and X.-S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM TIST*, vol. 2, no. 2, 2011, article no. 10.

[50] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 115–126, 2015.