Available online at www.sciencedirect.com
ScienceDirect
Biomedical Signal Processing and Control 1 (2006) 144–150
Biomedical Signal Processing and Control
www.elsevier.com/locate/bspc

Acoustic and perceptual assessment of vibrato quality of singing students

Noam Amir *, Orit Michaeli, Ofer Amir

Department of Communication Disorders, Sackler Faculty of Medicine, Chaim Sheba Medical Center, Tel-Aviv University, Tel-Hashomer, 52621, Tel-Aviv, Israel

Received 12 December 2005; received in revised form 15 June 2006; accepted 28 June 2006
Available online 7 September 2006

Abstract

While most studies that attempted to evaluate vibrato quality examined vocal productions of accomplished singers, very little is known about the characteristics of vibrato among singing students. Therefore, in this study, we performed a preliminary attempt to assess vibrato quality in their production of sustained notes. To that end, the presence and quality of vibrato in 253 sung notes was rated subjectively by five experienced singing teachers. The pitch contour was calculated for each recording, from which we calculated the FFT and the autocorrelation of this contour. Subsequently, a series of features was extracted from these two, and then different statistical methods were applied to examine whether the acoustic features could be used to define predictors that would be in agreement with the perceptual judgments. Given the moderate agreement obtained among judges, these acoustic predictors performed relatively well: vibrato existence was predicted correctly in over 82% of the recordings. The predictor for vibrato quality accounted for 46.5% of the variance of the subjective evaluation of vibrato quality. Due to the novelty of this study in assessing vibrato among students rather than among professional singers, several considerations and limitations, as well as directions for further research are discussed.

Keywords: Vibrato; Singing; Periodicity

1. Introduction

Vibrato in the singing voice has been the subject of several acoustical and perceptual studies, dating to the beginning of the 20th century. Recent studies of vibrato have focused on a detailed analysis of vibrato parameters, most often on productions of accomplished singers. For example, Prame [1] studied the variability in vibrato rates across 10 prominent artists, singing in the western classical music tradition. Analyzing 194 tones, he found an average rate of 6.0 Hz, with a mean variation from the singer's average of 8%. Prame also reported that most singers tended to accelerate the vibrato rate on the last five cycles of their vibrato. In a later study, Prame [2] examined vibrato extent and intonation over the same recordings. He found a mean extent of ±71 cents, with a negative correlation to tone duration. A study by Bretos and Sundberg [3] reexamined these parameters: rate, extent and intonation, with the addition of sound level. These were performed on selections from recordings of Verdi's *Aida*. Two high pitched notes of a soprano aria were selected. In general, their findings were consistent with those of Prame, though they found a higher degree of variability between singers. They also noticed the characteristic rise in vibrato rate towards the end of a note. Some dependencies on sound level were found, and vibrato regularity was found to be greater for lower notes.

Two recent studies differ from those quoted above, in that they compared acoustic parameters to human perception of the vibrato quality. Diaz and Rothman [4] selected notes sung by eight professional singers, and subjected them to judgment by four experienced listeners, who rated them as *good/pleasing* versus *poor/disagreeable*. Only notes that were judged unanimously as either good or poor were retained for acoustic analyses. Diaz and Rothman used an LPC based method for feature extraction from the pitch contour, coming up with two measures for rate variability and two measures for extent variability. Interestingly, they measured local rate and extent over segments of 170 ms, advancing in steps of 12 ms. This provided a measure to quantify how rate and extent varied over

* Corresponding author. Tel.: +972 3 5349817x109; fax: +972 3 5352868.
E-mail address: noama@post.tau.ac.il (N. Amir).

1746-8094/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bspc.2006.06.002

time. However, they included in their statistical analysis only the averages and S.D. of these measures and obtained significant results for the measures of extent variability. This could be interpreted to show that a constant vibrato extent is necessary for vibrato to be considered ‘‘good’’.

Another study, by Howes et al. [5], which compared acoustic measurements of vibrato to perceptual evaluation, also examined operatic singing. Extracts from arias in Donizetti's *Lucia de Lammermoor*, sung by five world famous singers were used for one part of their study. For the other part, an emotional cadenza from a soprano aria, taken from Verdi's *Un ballo in maschera*, was selected. This was also taken from recordings of renowned singers. Howes et al. performed several types of subjective judgments, from assessment of vibrato onset to assessment of emotional content. Their acoustic measurements, however, involved merely vibrato onset time, average rate and average extent. Evaluation of this portion of their results is difficult, because they make scant mention of whether certain trends in their results were statistically significant or not. Interestingly, however, they reported that vibrato rates across all singers were similar (6.28–7.59 Hz). They also concluded that the perception of vibrato quality did not exhibit a clear relationship with the measured acoustic features.

As noted, all published studies of vibrato that could be found in a literature search, were conducted on voices of accomplished singers. Despite this, these vibrato segments were not always judged to be of the highest quality. Nonetheless, the degree of uniformity to be expected amongst such singers is higher than what can be expected among singing students with varying degrees of singing experience.

In a preliminary study [6], we examined the influence of vocal warm-up on various measures of voice quality, among young classical singers. While there are widely accepted acoustic measures that are considered to be related to voice quality, at least for clinical applications, it was difficult to identify such consensual measures of vibrato quality. Our attempts to quantify the quality of the vibrato, and to correlate it with vocal warm-up, singing experience and perceptual ratings, led to the current study.

The singers in our preliminary study were all singing students, with a large variability in singing experience. Their control of vibrato was, at times, limited. Therefore, vibrato quality was found to vary greatly, from non-existent to very good. Based on the variability in vibrato quality we encountered among young singers, the aim of the present study was to identify acoustic measures that would correlate with the presence of vibrato and with its subjective evaluation. Though arguably such measures might not be applicable for studying nuances of vibrato across singing styles of professional singers, they could be useful for training of young singers. Such analysis, if successful, could ultimately lead to the development of bio-feedback type equipment which could provide singers with an assessment of their vibrato quality in real-time, or to serve as a helpful device for teaching singers. Furthermore, it is generally accepted that vibrato appears only after sexual maturation is completed, but its course of development has not previously been described. The study of vibrato as it develops through the stages of singers' education, as opposed to vibrato in accomplished singers, can shed light on the developmental traits of voice production from biomechanical, as well as psychosocial perspectives.

From a signal processing perspective, many of the recently performed studies performed F0 extraction through visual inspection of spectrograms [1,5]. The availability of computer programs such as Praat (http://www.praat.org), CSL (Kay Elemetrics) and others, enables accurate automatic measurement of F0, with smaller margin for human error. While the definitive evaluation of vibrato *quality* is based on subjective judgment, the related acoustic analyses can be performed nowadays using accurate and robust techniques. This has already been initially explored by Diaz and Rothman [4], who applied parametric modeling techniques to analysis of the F0 contour, though they did not specify how F0 was actually calculated. In the present study, therefore, we employ measurement methods that rely, to the largest extent possible, on automatic analysis, both of F0 and features extracted from the raw F0 contour.

2. Methods

2.1. Participants

Twenty young female singing students participated in this study, after obtaining the approval from our institutional review board and consent from all participants. All participants had professional classical voice training for a mean period of 5.4 years (S.D. = 2.9). Sixteen singers were conservatory students, and the other four were recently graduates of a music academy. Overall mean age was 18.62 years (S.D. = 3.2), mean weight was 61.5 kg (S.D. = 13.4) and mean height was 164.9 cm (S.D. = 6.1). All singers were healthy, with no remarkable medical history.

2.2. Recording procedure and instrumentation

Participants were recorded individually in a quiet room while sustaining the vowels /a/ and /i/ in three different pitches: 20, 50 and 80% of their reported vocal range. Each reference tone was presented by a piano in a random order, and the singer was asked to sustain the produced vowels (target tones) as accurately as possible for 3–5 s. The singers were not specifically instructed to produce vibrato in their sung notes. All vocal productions were recorded through a microphone (ACO Pacific Inc.) situated approximately 15 cm from the subject's mouth, using a Sony-TCD D7 digital recorder (Sony, Tokyo, Japan). Sampling rate for the recording was set to 48 kHz (16 bits per sample). Vocal productions of duration shorter than 2 s were also excluded from the analysis, leaving 253 recordings that were analyzed in all.

2.3. Subjective evaluation

The 253 recordings that were chosen for this study were presented, in a random order, to five experienced judges for evaluation. All judges were singing teachers, aged 30–58 with teaching experience ranging from 3 to 30 years. The judges were presented with a simple computerized questionnaire. For each recording, the judges were required to decide, first, whether it contained vibrato or not. If a recording was judged to contain vibrato, the judge was asked to rate its quality on a four-point scale, where 1 represents ''poor'', 2 ''fair'', 3 ''good'', and 4 ''very good''. The judges were allowed to listen to each recording only once, yet they could advance through the recordings at their own pace. Recordings that were judged by at least four judges as containing vibrato were considered, for the purpose of this study, as containing vibrato. The individual listening tasks lasted between 30 and 40 min.

2.4. Acoustic analysis

The acoustic analysis was performed based on the fact that vibrato is essentially, a periodic modulation in fundamental frequency, which is mostly found to be closely sinusoidal (e.g. [1,2]). Thus, the analysis must be based on a reliable measure of fundamental frequency. As mentioned above, this can be measured relatively easily using computer software. In the present study, we chose to implement a pitch detection algorithm (PDA) in Matlab, based on the autocorrelation method. This approach for extracting fundamental frequency has been discussed in numerous studies (e.g. [7]).

The original recorded productions varied in length between 1.5 and 5 s. Since the vibrato was audibly inconsistent in some of the recordings, an additional degree of uniformity was introduced to the perceptual and acoustic analyses by using only the last 2 s of each recording. As mentioned above, recordings shorter than 2 s were discarded.

F0 detection was performed over successive 20 ms windows (weighted with a Hamming window function), with overlap of 10 ms. The autocorrelation function over each window was interpolated by a factor of 16, giving temporal resolution of 1.42 microseconds in the detection of the pitch period. The resolution of the peak determination in Hz is thus dependent on frequency. The lowest resolution is obtained at the highest frequency (which has the smallest period). In our case this was 887.5 Hz, where the corresponding pitch resolution was 0.12 Hz (0.23 cents), which is perceptually negligible. Analysis of the last 2 s of each recording, using the above windowing scheme, resulted in 200 pitch points for each file. To verify the accuracy and validity of the pitch detection algorithm, the computed pitch contour for each file was inspected visually and corrected where necessary. Characteristic errors of PDAs are of the form of pitch halving or doubling, which are easily identified by visual inspection.

Most previous studies performed relatively basic analyses on the raw data, usually measuring vibrato rate and extent. Evidently, when studying the vibrato of professional singers, the vibrato is steady enough for these to be the dominant factors in determining its quality. In contrast, in a preliminary study [6], we found these features to be insufficient, and in some cases even inapplicable. The large variability in pitch contours produced by the students examined here required the use of

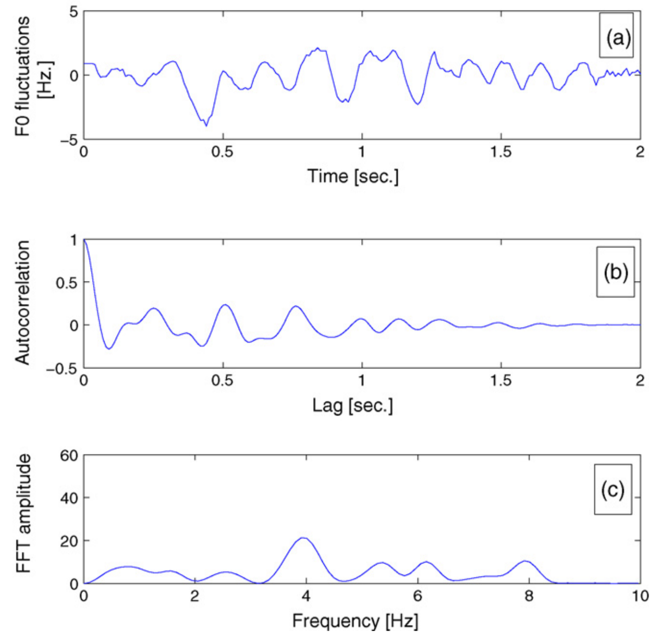

Fig. 1. A recording judged as having no vibrato: (a) F0 fluctuations; (b) autocorrelation; (c) FFT amplitude.

more general measures that would be able to detect whether vibrato exists at all, and assess its quality if present. To this end we employed methods originally used for detecting F0. Because F0 is defined as periodic oscillation in the voice signal, periodic oscillation of F0 (vibrato) can be measured using the same methods as well. We therefore applied two further analyses to the F0 contour: autocorrelation (after removal of DC) and the Fourier transform. Several illustrative examples are provided in Figs. 1–3. Fig. 1 demonstrates a

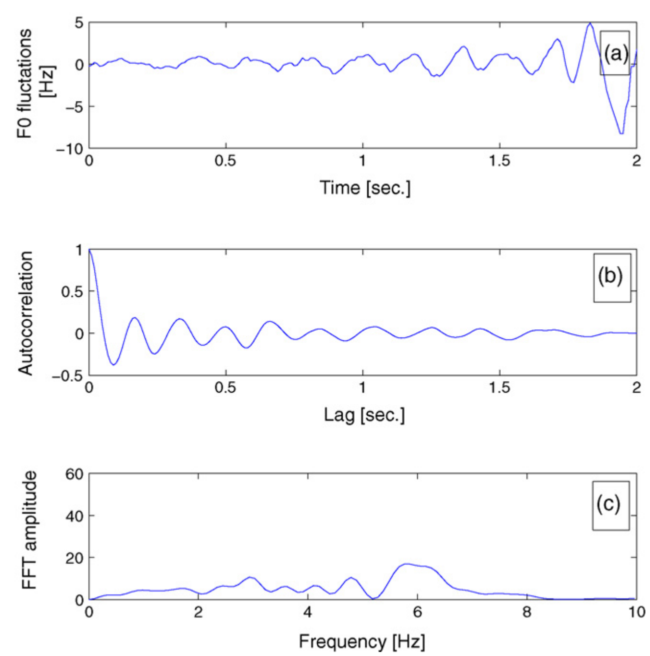

Fig. 2. A recording judged at an average vibrato rating of 1.4: (a) F0 fluctuations; (b) autocorrelation; (c) FFT amplitude.

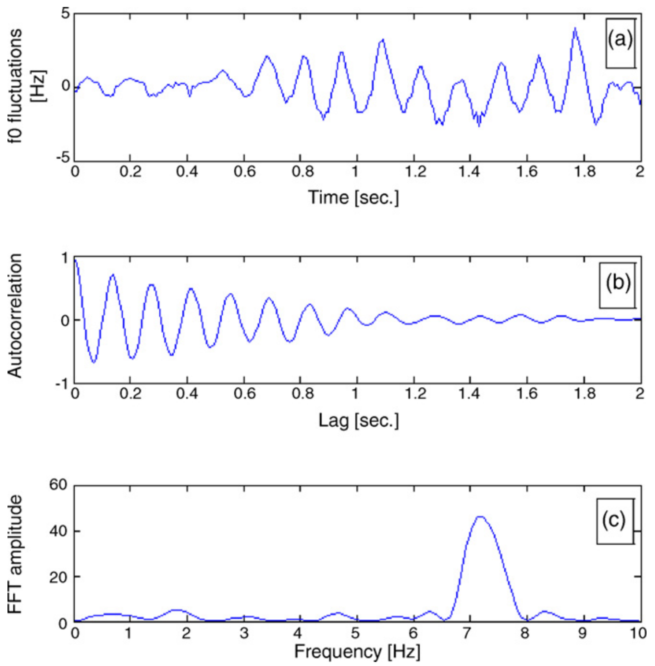

Fig. 3. A recording at an average vibrato rating of 2.8: (a) F0 fluctuations; (b) autocorrelation; (c) FFT amplitude.

recording, which was rated by the listeners, as not containing vibrato, Fig. 2 demonstrates an unsteady vibrato, which was rated by the listeners at 1.4, and Fig. 3 demonstrates an example of a steady vibrato, rated at 2.8. Each figure includes: (a) the F0 contour, after average has been removed, (b) the autocorrelation of the F0 contour, and (c) the Fourier transform of the F0 contour.

Visual inspection of these figures shows primarily that in the case of good vibrato, as shown in Fig. 3, there is a prominent single peak in the spectrum and a corresponding well-defined first peak in the autocorrelation function. In the case of lack of vibrato, as shown in Fig. 1, the pitch contour is chaotic, the first peak in the autocorrelation function is very low, and the spectrum is also rather chaotic. The intermediate cases, as exemplified in Fig. 2, on the other hand, are more difficult to interpret visually. We therefore treated the autocorrelation and the spectrum contours as raw data, from which meaningful features needed to be extracted.

Conducting a close visual examination of the autocorrelation and spectrum for many different productions, we observed that in many cases the indications of periodicity were in fact not manifested as a single prominent peak in the F0 spectrum. We often found concentrations of energy across the band of 5–7 Hz, though sometimes as more than a single peak. Also, the existence of a prominent peak in the autocorrelation contour, at lag lengths corresponding to these frequencies, seemed to provide a good indication of periodicity. This is not surprising, since autocorrelation is often used to measure F0 itself, as mentioned above.

In order to validate these observations and evaluate them quantitatively, we defined a series of measures taken from the above curves, with the objective of using statistical analysis to compare them to the subjective evaluations. These measures were based on the qualitative observations described above, with the aim of capturing the essence of these observations as closely as possible. The features were:

1. Energy between 4.5 and 7.5 Hz as compared to energy between 1 and 10 Hz.
2. Energy between 5 and 7 Hz as compared to energy between 1 and 10 Hz.
3. Location and height of the first peak in the autocorrelation of the F0 contour.
4. Location and height of the first trough in the autocorrelation.
5. Variance of the F0 contour.
6. Local extent of vibrato was measured by taking the range of the F0 values over a sliding window on the F0 contour, of 10 F0 points (100 ms). Mean and standard deviation of the resultant curve were taken.
7. Frequency of highest peak above 2 Hz in the FFT contour.
8. Height of highest peak above 2 Hz in the FFT contour.
9. Quality factor of highest peak above 2 Hz in the FFT contour.
10. Number of spectral peaks above half/third/quarter of highest peak above 2 Hz in the FFT contour.

3. Results

The analysis was carried out in several stages. First we evaluated the agreement between judges, since these subjective judgments were the baseline to which the acoustic measures were later compared. This comparison was carried out on two levels: the first, to determine a measure for presence/absence of vibrato, and the second, to identify a measure that correlates with the perceptual judgment of vibrato.

3.1. Agreement among judges

To assess agreement among judges, ratings indicating absence of vibrato were assigned a numerical value of 0, while presence of vibrato was signified by the rating of 1–4. Agreement among judges was first assessed using Kendall's coefficient of concordance [13], and yielded a value of 0.524 ($p < 0.001$). This is typically considered a moderate value, which can be interpreted to show that the listening task was relatively challenging, and that judges had somewhat different internal scales for assessing vibrato quality. This was further corroborated by calculating Cohen's κ [14], which yielded a score of 0.142 for the overall ratings. This analysis provided more revealing information, though, when the variation of these

Table 1
Cohen's κ for the different subjective ratings

Judgment	Cohen's κ
0	0.289
1	0.084
2	0.006
3	0.150
4	0.130

scores was observed among the different ratings (see Table 1). Data show that the judges were in best agreement at the extremes of the scale: Cohen's κ is highest for judgments of 0 (lack of vibrato) and for ratings of 3 and 4 ("good" and "very good").

The judges were also requested to provide a short verbal description of their impressions during the listening test. Overall, the judges commented that the task was indeed challenging. Some of them found it difficult to separate their impression of vibrato from other aspects of voice quality, and to distinguish between lack of vibrato and poor vibrato. Two judges suggested that, perceptually, vibrato was better in the higher registers.

3.2. Performance of acoustic measures

3.2.1. Presence of vibrato

The subjective judgments were degenerated into a binary variable in this phase. Recordings that were rated by four or five judges as containing vibrato received a value of "yes" (129 out of the 253 recordings), and recordings that were rated by three judges or fewer to contain vibrato, received a value of "no" (124 out of 253). Logistic regression [15] was then applied to the raw measures presented in the previous section, in order to find which of them were in best correlation with the subjective judgments, and to define a predictor that would be in optimal agreement with them. As expected, some of the acoustic measures were highly correlated with each other, for example measures 1 (energy between 4.5 and 7.5 Hz) and 2 (energy between 5 and 7 Hz), which yielded a Pearson correlation coefficient of $r = 0.89$ ($p < 0.0001$). Instead of manually choosing the most salient features, an automatic forward selection process was carried out, in which successive measures were added to the list until negligible improvement was obtained. The process is summarized in Table 2.

Though some of the features are highly correlated, as demonstrated above, they provide an overall result that gives good agreement with the judges' ratings. This is borne out by the logistic regression c statistic (area under the ROC curve), of $r = 0.904$.

Using the results to define a predictor of vibrato presence/absence, we obtained the results summarized in Table 3.

Table 3 shows an overall recognition rate of 82%. In addition, false negatives are slightly more prevalent (20%) than false positives (16%).

Table 2
Summary of features (in order of selection) in the forward selection processs

Feature (in order of selection)	Odds ratio
Energy between 4.5 and 7.5 Hz	41.7
Height of first autocorrelation trough	33.9
Height of first autocorrelation peak	53.0
Mean of local extent	10.3
Variance of F0 contour	1.8
Energy between 5 and 7 Hz	126.4
Lag of first autocorrelation trough	1.26

Table 3
Classification results for vibrato existence

Actual	Predicted		
	No	Yes	Total
No	104	20	124
Yes	26	103	129
Total	130	123	253

3.2.2. Rating of vibrato

The acoustic measures were analyzed statistically in order to find a predictor that would correlate well with the judges' average rating of the recordings judged to contain vibrato. Using a linear regression analysis with a forward selection process, a total of 12 features explained 45.6% of the variance (equivalent to a predictor which had a correlation of 0.675 with the judges' average ratings), which is considered a moderate value. Nonetheless, a reduced set of only five features explained 40.8% of the variance. In order of importance, these were:

1. absolute height of highest peak above 2 Hz in the FFT of the pitch contour,
2. energy between 4.5 and 7.5 Hz,
3. location of first autocorrelation peak,
4. energy of main spectral peak divided by its width,
5. the number of spectral peaks above one quarter of the highest peak.

4. Discussion

4.1. Subjective judgments

Interjudge agreement, in the perceptual task in this study, was moderate, as reflected by the value of 0.524 for Kendall's coefficient of concordance. This result was attained despite the special attention that was given to the selection of judges, in an attempt to improve interjudge agreement. The five judges who participated in this study were highly qualified and well-trained singing teachers. They were all accustomed to listening to voices of singing students. In addition, they were informed, prior to performing the listening task, that the recordings were taken from students rather than from accomplished singers. Thus, our results demonstrate that obtaining a high level of agreement on vibrato assessment among judges is difficult to expect. Nevertheless, agreement level on *existence* of vibrato (a dichotomic task) was higher than agreement level on judgment of vibrato *quality*.

Several possible explanations for the moderate agreement in the listeners' judgment could be considered. First, perceptual tasks of assessing voice quality were previously shown to have relatively low interjudge reliability. Such findings were reported in relation to judgments of specific perceptual vocal features, such as strain and asthenia [8,9]. This issue was, recently, addressed by Shrivastav [10] who suggested that improving interjudge reliability in perceptual listening tasks could be achieved by having the listeners perform multiple judgments of each token, and then averaging these. It is possible, therefore, that by modifying the listening task paradigm, higher levels of interjudge agreement could be obtained. Nonetheless, this would be at the expense of a longer listening test, which might present problems of listener fatigue or impatience. It should be noted that vibrato is usually used for expressive purposes by proficient singers, and judges would be expected to relate to it as such. The judges in the present study, however, were required to evaluate vibrato in isolated phonation, and not within a musical phrase. Thus their experience in assessing vibrato may not have been exploited to its fullest capability.

An alternative explanation could be attributed to the fact that the listeners in this study were required to provide a global judgment of vibrato, based on two seconds of singing. A recorded voice sample of two seconds, taken from a professional singer, might be sufficient for identifying and judging vibrato existence and quality. However, such a recording, taken from a singing student, could be more difficult to evaluate. Inspection of the recordings used in this study revealed that vibrato quality, as well as its existence, was not always consistent across the two seconds of the recording. Therefore, it is possible that different listeners reacted to different segments of the recordings during the listening task. Further support for this hypothesis can be found in the fact that listeners' agreement on vibrato quality was significantly better on the extremities of the rating scale (i.e., when ratings were either markedly high or low). It is possible that judging vibrato of high quality is similar to judging vibrato of accomplished singers, and therefore listening agreement was better. In addition, identifying lack of vibrato, or vibrato of extremely low quality is more of a dichotomic task, and thus it might be an easier task for the judges. Therefore, judging vibrato of moderate quality could be more difficult than extremely high or low quality vibrato. In order to evaluate this possibility, it might be necessary to construct a more localized listening task, in which judges could listen to longer recordings, and segment them according to vibrato quality. Such a paradigm could provide further insight into the issue of the required segment duration for a reliable identification of vibrato existence and quality.

Finally, a third explanation for the relatively low interjudge reliability could be attributed to the influence of additional voice properties, to which the singing teachers might have reacted during the listening task. Although the judges were instructed to respond strictly to the vibrato quality, singing teachers are more apt to judge a singing segment using a holistic perspective. Thus, judging vibrato quality while overlooking other voice properties, such as vocal stability, intensity or timbre might be difficult, and could require special practice. Furthermore, it was previously shown that the judgment of vibrato quality is strongly dependent on personal singing style and on individual preferences of the listener [5]. Therefore, it is possible that although the judges in this study were professional singing teachers, their judgments were influenced by their personal musical or singing style preferences.

4.2. Acoustic analysis

The results for vibrato *existence* shown above revealed 82% agreement between the judges' decision and the automated decision based on the acoustic features. This demonstrates that good agreement can be obtained between the subjective judgments and an automated predictor. Though a list of seven acoustic features were used by this predictor, three of them were the most dominant: (a) energy in the 4.5–7.5 Hz band, (b) the height of the first autocorrelation trough, and (c) the height of the first autocorrelation peak. Interestingly, the height of the first autocorrelation peak is very often used as an indicator of periodicity in the speech signal itself, therefore it serves as the basis for various pitch detection algorithms. The first autocorrelation trough, on the other hand, is not a widely used feature. The decision to include this feature in our analysis was based on visual inspection of the autocorrelation curves of our data, where we identified a common pattern: clear cases of periodicity were also characterized by a marked first peak and trough in the autocorrelation curve. This indicates that the first trough carries valuable information, which is independent from that conveyed by the first peak. This was verified by calculating the Pearson correlation coefficient between them, yielding a low correlation value ($r = 0.368$, $p < 0.05$). A possible conclusion that can be drawn from the fact that these three measures were the best predictors of vibrato is that vibrato is most likely to be identified when there is a clear periodicity in the F0 contour, providing that this periodicity is between 4.5 and 7.5 Hz.

The results for rating of vibrato *quality* are not as conclusive as the results for vibrato existence. The correlation between the subjective and objective measures was moderate ($r = 0.675$), and only 46.5% of the variance could be explained by these measures. This result should be evaluated in light of the relatively low agreement among judges. It is expected that if interjudge agreement would be improved, a better correlation between perceptual and acoustic results could be obtained. For example, identifying the optimal interval duration for the perceptual task could increase interjudge agreement, providing a more solid basis for the comparison to the acoustic analysis paradigm. Furthermore, the present results of improved agreement at the extremities of the scale are encouraging since these segments of the scale are of most interest, from a singing education perspective.

Similar to results found for vibrato existence, a limited set of features was the most crucial for rating of vibrato *quality*. The three most prominent features were (a) absolute height of the first spectral peak above 2 Hz, (b) energy between 4.5 and 7.5 Hz, and (c) the location of the first autocorrelation peak. Interestingly, a small number of features was the most significant in explaining both existence and rating, but out of the three most important features for each case, only one was common to both. This can be interpreted to suggest that the two tasks should indeed be performed and analyzed separately. This is reminiscent of the separation between voiced/unvoiced decisions in the speech signal, and measuring F0 where voicing is found to be present.

The results obtained in the objective part of this study could be examined from two alternative points of view. On the one hand, it is interesting to examine the way in which these results can be used for providing a singing student with feedback about his or her vibrato quality. For this purpose, the number of features included in the model is of little consequence, and a more liberal approach can be adopted with respect to the fact that some of the features might be intercorrelated, while others might only have a small contribution to the outcome. In this case, our sole concern is to obtain a predictor with the highest possible correlation with the subjective judgments. On the other hand, from a theoretical perspective, we are interested in identifying the smallest set of features that would correlate best with the listeners' perception of vibrato. We elaborate on this in the following section.

4.3. Dominant acoustic features

Several parallels can be drawn between the evaluation of vibrato and the evaluation of voice quality in sustained phonation. Previous research has suggested that specific acoustic measures, which are associated with voice quality, such as jitter and shimmer, are applicable only when periodicity or quasi-periodicity is present in the *voice signal* [11,12]. When these perturbation values deviate extremely from the norm, this could be an indication of poor periodicity, which undermines the validity of these measures. Similarly, most studies of vibrato, which were performed on professional singers, examine cases of clearly defined periodicity in the *pitch contour*. Therefore, these studies could base their analyses on changes observed between consecutive cycles. In the present study, however, a different set of features needed to be considered, because periodicity in the F0 contour (i.e. steady vibrato) could not be taken as a preliminary assumption. Hence, determining a set of successive periods was not always possible.

The features we selected for the analysis were based on concentration of energy in a narrow part of the spectrum, on the one hand, and indications of periodicity found in the autocorrelation contour, on the other hand. Interestingly, important features were found, independently, in the autocorrelation contour as well as in the FFT contour. In effect, the two analyses were complementary, since it is well known from transform theory that features localized in one domain are spread out in the other. Although initially we experimented with a relatively large set of features, two main sets of features emerged as the dominant ones. In the FFT contour, the most important features were those that measured the localization of energy in the approximate band of 5–7 Hz. In the recordings that were found to contain vibrato, its quality was judged better when the energy was further concentrated, and presented a single high and narrow peak. This indicates that high quality vibrato is identified when it is stable and nearly sinusoidal. Otherwise, further harmonics of the fundamental vibrato frequency would become significant. Further, in the autocorrelation contour, the most prominent features were those associated with the first peak and trough. As noted before, this combination of a clear negative trough followed by a peak is usually clear indication of periodicity.

In conclusion, the traditional features used to assess vibrato in previous studies were found to be appropriate only when clear periodicity in the F0 contour (i.e. vibrato) is present. The current study defines a set of features that could be applicable for identifying presence of vibrato and for assessing its quality, when clear periodicity is not evident. The selected set of acoustic features, which was based on FFT and autocorrelation analyses, yielded encouraging though preliminary results. Further research, based on a wider range of vocal features, such as modulations of amplitude and formant values, is needed to enhance our understanding of the relationship between subjective and instrumental evaluation of vibrato.

Acknowledgment

The authors would like to thank Ms. Esther Shabtai for her advice and assistance in the statistical analysis of the data.

References

[1] E. Prame, Measurements of the vibrato rate of 10 singers, J. Acoust. Soc. Am. 96 (4) (1994) 1979–1984.
[2] E. Prame, Vibrato extent and intonation in professional Western lyric singing, J. Acoust. Soc. Am. 102 (1) (1997) 616–621.
[3] J. Bretos, J. Sundberg, Measurements of vibrato parameters in long sustained crescendo notes as sung by 10 sopranos, J. Voice 17 (3) (2003) 343–352.
[4] J.A. Diaz, H.B. Rothman, Acoustical comparison between samples of good and poor vibrato in singers, J. Voice 17 (2) (2003) 179–184.
[5] P. Howes, J. Callaghan, P. Davis, D. Kenny, W. Thorpe, The relationship between measured vibrato characteristics and perception in western operatic singing, J. Voice 18 (2) (2004) 216–230.
[6] O. Amir, N. Amir, O. Michaeli, Evaluating the influence of warm-up on singing voice quality using acoustic measures, J. Voice 19 (2) (2005) 252–260.
[7] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, IFA Proc. 17, 97–110.
[8] M.S. De Bodt, F.L. Wuyts, P.H. Van de Heyning, C. Croux, Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality, J. Voice 11 (1) (1997) 74–80.
[9] H. Yamaguchi, R. Shrivastav, M.L. Andrews, S. Niimi, A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale, Folia Phoniatr. Logo. 55 (3) (2003) 147–157.
[10] R. Shrivastav, C.M. Sapienza, V. Nandur, Application of psychometric theory to he measurement of voice quality using rating scales, J. Speech Lang. Hear Res. 48 (2) (2005) 323–335.
[11] I.R. Titze, Workshop on Acoustic Voice Analysis: Summary Statement, National Center for Voice and Speech, Iowa City, 2005.
[12] E.P. Ma, E.M. Yiu, Suitability of acoustic perturbation measures in analysing periodic and nearly periodic voice signals, Folia Phoniatr. Logo. 57 (2005) 38–47.
[13] M.G. Kendall, Rank Correlation Methods, 2nd ed., Hafner Publishing Co., New York, 1955.
[14] J.L. Fleiss, Statistical Methods for Rates and Proportions, 2nd ed., John Wiley and Sons, New York, 1981.
[15] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, John Wiley and Sons, New York, 2000.