ROBUST FEATURE EXTRACTION FOR AUTOMATIC RECOGNITION OF VIBRATO SINGING IN RECORDED POLYPHONIC MUSIC

Felix Weninger[1]*, Noam Amir*[2]*, Ofer Amir*[2]*, Irit Ronen*[2]*, Florian Eyben*[1]*, Björn Schuller*[1]

[1] Institute for Human-Machine Communication, Technische Universität München, Germany
[2] Department of Communication Disorders, Tel Aviv University, Israel

weninger@tum.de

ABSTRACT

We address the robustness of features for fully automatic recognition of vibrato, which is usually defined as a periodic oscillation of the pitch (F0) of the singing voice, in recorded polyphonic music. Using an evaluation database covering jazz, pop and opera music, we show that the extraction of pitch is challenging in the presence of instrumental accompaniment, leading to unsatisfactory classification accuracy (61.1 %) if only the F0 frequency spectrum is used as features. To alleviate, we investigate alternative functionals of F0, alternative low-level features besides F0, and extraction of vocals by monaural source separation. Finally, we propose to use inter-quartile ranges of F0 delta regression coefficients as features which are highly robust against pitch extraction errors, reaching up to 86.9 % accuracy in real-life conditions without any signal enhancement.

Index Terms— Singing style, music signal processing, feature extraction

1. INTRODUCTION

Vibrato singing is usually characterized as a periodic oscillation of the fundamental frequency (pitch) of the voice at a rate of 4–8 Hz. Applications of automatic recognition of vibrato singing in recorded polyphonic music include singer identification, as different singers develop their own style of vibrato [1], as well as other music information retrieval tasks such as structure and performance analysis. Furthermore, it can be useful for highly efficient audio coding, e. g., as an attribute for sound synthesis [2].

Few performance studies exist on fully automatic recognition of vibrato singing [3, 4]; furthermore, these are limited to monophonic recordings, which may be justified for applications such as coaching of singing students. For retrieval applications in recorded music, however, one has to cope with additional sources from accompaniment. Per definition, it can be assumed that automatic recognition of vibrato strongly depends on robustness of pitch extraction, which, however, is challenging in the condition of multi-source mono- or stereophonic recordings [5].

In this study, we use an evaluation database of polyphonic music covering different musical genres (jazz, pop and opera) to investigate the effect of 'noise' by instrumental accompaniment on pitch extraction and resulting classification accuracy of vibrato. Besides the obvious frequency analysis of the F0 contour [2], we consider other functionals of pitch itself, and besides, the contours of other 'low-level' features. Finally, we employ extraction of the leading voice by unsupervised source separation to contribute to robustness of pitch extraction. 'Unsupervised' in this respect means that no a-priori knowledge about the nature of the accompaniment, or the voice of the singer, is assumed [6].

The research leading to these results has been partly supported by the German Research Foundation (DFG) through grant no. SCHU 2508/2.

Genre	Non-vibrato	Vibrato	Sum
Jazz	34	96	130
Pop	50	90	140
Opera	–	160	160
Sum	84	346	430

Table 1: Number of instances per class (non-vibrato / vibrato) and genre (jazz / pop / opera) in the evaluation database.

The remainder of this contribution is structured as follows: Our evaluation database is described in Section 2. We move on to outline our feature extraction approach in Section 3 and examine the results of automatic classification experiments in Section 4. A statistical perspective on the relevance of various functionals of the pitch contour is given in Section 5 before concluding in Section 6.

2. EVALUATION DATABASE

Our evaluation database consists of 430 segments, each corresponding to one sung note in professionally recorded music. 30 different artists are found in the database, all of which are accomplished female singers. Genres cover jazz, pop and opera with 10 singers each, and 130, 140 and 160 instances, respectively. The segments were extracted manually and labeled by experts as containing vibrato or not. All opera segments are sung with vibrato while the percentage of vibrato segments is approximately 2/3 for pop and 3/4 for jazz. The average note duration is 1.86 s with a standard deviation of 1.10 s and considerably differs among genres, with jazz exhibiting at the same time the longest average duration (2.12 s) and the greatest standard deviation (1.54 s). In this study, unlike in [4], we do not pre-select notes by their duration; further, to increase realism, we intentionally include instances where vibrato is delayed in the note, as is often the case in jazz: In our database, a delay occurs for 52 of 96 instances (54 %), reaching up to 81 % of the total note duration.

In this study, we consider it as crucial to force a singer-independent subdivision for automatic classification experiments, as vibrato styles differ considerably among singers [1], and a singer-dependent system would be prone to over-fitting to the specificities of the singers in the database. Thus, we subdivided the database into three folds for singer-independent cross-validation stratified by genre, i. e., the class and genre distributions in each fold are approximately equal. For the sake of reproducibility, these folds were obtained as follows: For each genre, the ten singers were sorted alphabetically and assigned to fold 1 (singers 1–4), fold 2 (singers 5–7)

LLD	Functionals
(Δ) Log. F0	*Extremes*
(Δ) RMS energy	Range
(Δ) Auditory spectrum	Position min., max.
	Dist. min./max. from arith. mean
	Moments
	Std. dev., skewness, kurtosis
	Temporal evolution
	Mean crossing rate (MCR)
	DCT coefficients 1–6
	Percentiles
	Inter-quartile ranges 1–2, 2–3, 1–3
	DFT coefficients 1–10 (log. F0)
	Arith. mean (of Δ)

Table 2: Feature extraction: Low-level descriptors (LLD) and functionals.

or fold 3 (singers 8–10). Consequently, of the 430 instances in the database, 170 / 130 / 130 are assigned to each of the three folds.

3. FEATURE EXTRACTION

3.1. Low-Devel Descriptors

As frame-wise low-level descriptors (LLDs), we extract pitch (F0), root mean square (RMS) energy and auditory spectrum. All feature extraction is performed without manual post-processing such as correction of octave errors. Pitch detection is based on the Subharmonic Summation (SHS) algorithm [7] to identify pitch candidates in the frequency domain. Thereby the power spectrum is transformed to a logarithmic scale by spline interpolation and shifted spectra are added to the original spectrum to sum up the harmonics. The result is the so called SHS spectrum (SHSS). In theory there should be one prominent peak at F_0; however, in practice higher harmonics are also present. The N highest peaks in the SHSS are identified, and peak position and amplitude are adjusted by three point quadratic regression using the peak and its left and right neighbors to fit a parabola. A voicing probability is assigned for each candidate based on the (adjusted) peak's amplitude in the SHSS. The arithmetic mean (μ_s) of the bins in the SHSS is computed. For each pitch candidate i with a pitch candidate score s_{ci} (= peak amplitude) greater than μ_s the voicing probability p_{vi} is computed as $p_{vi} = 1.0 - \frac{\mu_s}{s_{ci}}$. Otherwise ($s_{ci} \leq \mu_s$), $p_{vi} = 0$. The final pitch contour as well as the final voicing decision is smoothed by dynamic programming where soft penalties for jumps and out-of-range values are applied. The algorithm is based on the Viterbi pitch smoothing as presented in [5], which was slightly modified for the SHS pitch values and voicing probabilities. This implementation of pitch extraction is available in our open-source toolkit openSMILE [8].

To make the absolute amount of pitch variation independent of the fundamental frequency, we use the natural logarithm of the pitch. Pitch and RMS energy are extracted from 50 ms frames of the audio signal windowed with a Gaussian function at 10 ms frame shift. The auditory spectrum is computed by reweighting the Mel frequency bands 1–26 obtained from a short-time Fourier transform (STFT) with 25 ms frame size and a Hamming window function, similarly to the procedure performed in extraction of Perceptual Linear Prediction features.

3.2. Functionals

To capture variation of the low-level descriptors, first order delta regression coefficients (Δ)—a kind of discrete derivative often used in speech processing—are extracted according to [9] spanning 5 frames. Furthermore, segment-wise functionals are computed from both the low-level descriptors and their delta coefficients. To capture pitch oscillations in the range relevant for vibrato, Discrete Fourier Transform (DFT) coefficients 1–10 are extracted from overlapping windows of 128 logarithmic F0 points which are normalized ('centerized') to zero mean, corresponding to a window size of 1.28 s to achieve sufficient frequency resolution. Windows overlap by 64 points and are multiplied by a Hamming function before applying the Fourier transformation. Zero padding is applied for segments shorter than the length of a single window (1.28 s); for segments longer than a single window, incomplete windows at the end are discarded—otherwise, we often observed the DFT coefficients from the previous windows to be deteriorated by the alteration of the frequency distribution in the last window due to zero padding. Finally, the mean across windows is taken for each DFT coefficient.

Other, more generic functionals applied to all kinds of LLDs include moments, range (absolute difference of minimum and maximum), distance of minimum and maximum from the arithmetic mean, standard deviation and higher order moments, mean crossing rate, Discrete Cosine Transform (DCT) coefficients 1–6 and finally inter-quartile ranges (IQR, absolute differences between quartiles). These functionals are often used in segment-based functional extraction from pitch and other LLDs for paralinguistic information retrieval [10]. Frames (erroneously) classified as unvoiced are excluded from calculation of functionals and Δ coefficients from the F0 contour, except for DFT coefficients to preserve the frequency of periodic oscillations—in that case, unvoiced frames are assumed to be equivalent to the mean of the F0 points in the corresponding window(s). Note that we choose only functionals which are independent of the absolute values of the LLDs in order to capture signal variation instead of overall characteristics. Thus, for instance, the arithmetic mean is only computed from the delta coefficients, not from the LLDs themselves.

4. AUTOMATIC CLASSIFICATION EXPERIMENTS

4.1. Preprocessing

As an unsupervised preprocessing method for extraction of the singer's voice in the presence of background music, we use the leading voice separation approach described in [6]. This approach differs from traditional pitch extraction by explicit modeling of accompaniment and the vocal tract of the singer. More precisely, the STFT of the observed signal at each frame is expressed as the sum of STFTs of vocal and background music signals. These are estimated by an unsupervised approach building on non-negative matrix factorization techniques: The voice STFT is modeled as product of source (periodic glottal pulse) and filter STFTs while no specific constraints are set for the background music signal because of its wide possible variability. The estimation of the various model parameters is then conducted by iterative approaches. In particular, the initial parameter estimate is refined by Viterbi smoothing in order to limit, e. g., octave jumps of the voice. To ensure best reproducibility of our findings, we used an open-source implementation[1] of the algorithm with default parameters.

[1] Software available at http://www.durrieu.ch/phd/software.html

To get an upper performance bound by simulating near-perfect pitch extraction, we apply a band-pass filter for each segment whose pass-band was manually set to capture single harmonic(s) of the singing voice including pitch variations, so that robust automatic pitch determination is straightforward. These filters were applied to the Fourier transform of each segment as a whole to achieve best frequency resolution.

4.2. Classification

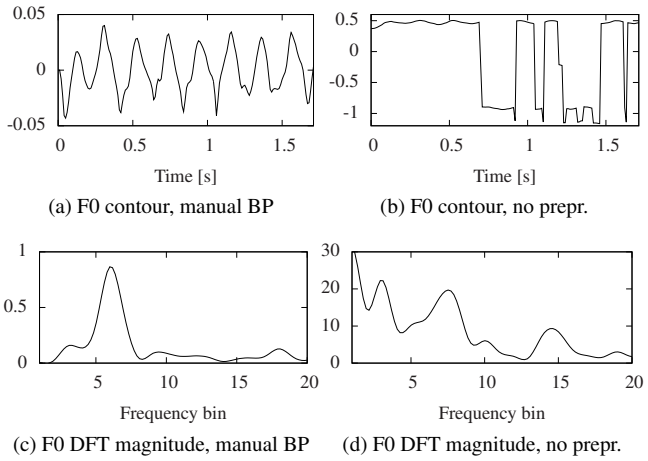
For classification we used the SimpleLogistic algorithm [11] implemented in the Weka toolkit [12]. This classifier is particularly suitable for small to medium feature spaces as it is based on boosting of one-dimensional regression functions. The number of boosting iterations was cross-validated on the training set, using the default parameters in the Weka toolkit. In order to optimize on a balanced recall of both the vibrato and non-vibrato classes, we applied training instance up-sampling as follows: For testing on each fold, all instances of the minority class (non-vibrato) in the two folds used for training were copied so as to achieve uniform a-priori class probabilities in the training set for the classifier.

4.3. Results and Discussion

In Table 3, we present the unweighted average recall (UAR) of the vibrato and non-vibrato classes. This measure is arguably better suited to take into account the class imbalance in the database than is conventional accuracy. We observe that the joint feature space of F0 functionals provides highly robust classification, reaching 84.9 % UAR without preprocessing. Still, this is significantly ($p < .05$ according to a one-tailed z-test) below the upper bound of 90.1 % achieved by manual band-pass filtering, indicating that accompaniment makes vibrato classification considerably more challenging. Surprisingly, vocal separation significantly degrades performance by over 8 % absolute compared to using no signal enhancement; this can be attributed to the fact that the algorithm was not designed to capture slight pitch variations, as the source STFT is assumed to be constant in the vocal model [6].

An analysis of different functional groups of F0 reveals that as expected, DFT coefficients allow highly robust classification on their own if pitch extraction is facilitated by manual bandpass filtering (91.7 % UAR); this performance, however, is vastly degraded to 61.1 % UAR without such preprocessing. Notably, vocal separation can partly alleviate this downgrade (66.1 % UAR). An explanation for the performance drop is shown in Fig. 1 for an example of jazz music. It is clearly visible that while there is a prominent peak in the 6th frequency bin (4.7 Hz) for 'ideal' F0 extraction, in the real-life setting without manual preprocessing the DFT coefficients are significantly deteriorated due to the pitch estimation errors. In the example, these are caused by chords played by a piano in the second half of the segment.

As to other functionals, extremes and moments both are robust (around 80 % UAR) for bandpass filtering, yet performing not significantly ($p > .05$) above chance level (50 % UAR) without preprocessing, probably due to sensitivity against outliers. Their performance apparently cannot be restored by vocal separation. In contrast, DCT and MCR perform similarly above chance level regardless of the preprocessing, but generally deliver unsatisfactory accuracy on their own (up to 66.2 % UAR). Finally, and most importantly, percentile features, i. e., inter-quartile ranges enable highly robust classification (86.9 % UAR) even without preprocessing; for bandpass filtering, their UAR of 88.0 % is still remarkable, yet significantly below the one of DFT coefficients. This behavior will be further investigated below.



(a) F0 contour, manual BP (b) F0 contour, no prepr.

(c) F0 DFT magnitude, manual BP (d) F0 DFT magnitude, no prepr.

Fig. 1: Logarithmic F0 contour (normalized to zero mean) and its DFT coefficients for a segment of jazz music, without or with manual bandpass (BP) filtering to extract the singing voice.

Unweighted average recall [%]		Preprocessing		
LLD	Functionals	—	Voc.sep.	BP
(Δ) Log. F0	All	84.9	73.3	90.1
	DFT coeff.	61.1	66.1	**91.7**
	DCT / MCR	65.4	60.9	66.2
	Extremes	55.5	51.8	81.1
	Moments	56.4	53.5	79.1
	Percentiles	**86.9**	**78.0**	88.0
(Δ) Aud.spec.	All	67.0	59.7	61.9
(Δ) Energy	All	59.5	67.0	73.8
All	All	67.1	65.0	77.7
(Δ) F0 + Energy	All	79.8	75.1	82.7

Table 3: Unweighted average recall of vibrato / non-vibrato instances in singer-independent 3-fold stratified cross-validation using SimpleLogistic classification by feature set and functional groups. Preprocessing by vocal separation or manual band-pass (BP).

Before, we briefly summarize performance of other LLDs than pitch. While both, auditory spectrum and RMS energy cannot compete with F0 in terms of UAR, it is notable that both are observed highly above chance level ($p < .005$). Interestingly, the auditory spectrum seems to carry valuable information especially when using no preprocessing (67.0 % UAR); this is possibly due to the strong interdependency with musical genre (cf. Table 1), which can be detected by spectral features. Furthermore, energy is considerably informative; this is *not* simply due to vibrato occurring in accented notes of higher loudness since the functionals are independent of the absolute energy (see above). Still, none of the 'alternative' LLDs seem to complement the information gained from the pitch, as classification with the union of feature sets (all or F0 + energy) cannot significantly surpass the performance of F0 functionals alone.

Overall, the effect of vocal separation is disappointing for our task. One could possibly extend the algorithm to adapt the source model to slight variations; still, given the stability of the pitch features without signal preprocessing, a large performance increase would be required to outweigh the high additional computational effort from a practitioner's point of view.

5. FEATURE RELEVANCE

As a perspective on feature relevance independent of the classification algorithm, we perform two-sided Welch two-sample t-tests (assuming inequal variance) on the features derived from F0 and its delta regression coefficients in the whole data set. These tests indicate whether their mean in the vibrato segments is significantly different from the mean in the non-vibrato segments. We do not correct the p-values for repeated measurements since we are only interested in a feature ranking. Furthermore, we restrict this evaluation to F0 functionals due to their vastly superior performance in general (cf. the previous section).

For each of manual bandpass filtering as well as no signal enhancement, the ten most discriminative functionals of F0 and their delta regression coefficients by their absolute t-statistic are shown in Tables 4a and 4b, respectively. Evidently, inter-quartile ranges of Δ F0 are particulary informative in both cases. Inter-quartile ranges of F0 itself, however, are only informative for manual BP filtering. This indicates that deltas are robust against pitch estimation errors while F0 itself is not: Apparently, due to the Viterbi smoothing, the musical accompaniment causes systematic errors in pitch estimation rather than random fluctuations. Furthermore, it is well known that generally, percentile-based features such as IQR are robust against outliers caused by measurement errors. Concerning DFT coefficients, it is obvious that they have strong discriminative power after applying manual BP filtering: The most relevant coefficients 6–10 exactly correspond to vibrato rates from 4.7 to 7.8 Hz. In accordance with the automatic classification experiments, they are less discriminative when using no preprocessing; furthermore, although DFT coefficients 1 (of Δ F0), 4 and 5 occur in the 10 most relevant features, their relation to vibrato is not immediately obvious: Coefficient 1 corresponds to 0.77 Hz, and the t-statistic of coefficients 4 and 5 is negative. Thus, these functionals (as well as the first DCT coefficient and mean crossing rate) apparently provide an useful, yet generic assessment of temporal evolution.

LLD	Functional	p	t
Δ F0	IQR 1–2	$\ll .001$	12.9
Δ F0	IQR 1–3	$\ll .001$	12.8
Δ F0	IQR 2–3	$\ll .001$	10.7
Δ F0	DFT coeff. 1	$< .001$	4.2
F0	DCT coeff. 2	$< .001$	- 3.8
F0	Mean crossing rate	$< .001$	3.6
F0	DFT coeff. 5	$< .005$	- 3.2
Δ F0	Position of max.	$< .005$	- 3.2
Δ F0	Position of min.	$< .01$	- 2.8
F0	DFT coeff. 4	$< .01$	- 2.6

(a) no preprocessing

LLD	Functional	p	t
Δ F0	IQR 1–3	$\ll .001$	19.9
Δ F0	IQR 2–3	$\ll .001$	19.7
Δ F0	IQR 1–2	$\ll .001$	18.9
F0	DFT coeff. 7	$\ll .001$	13.2
F0	DFT coeff. 8	$\ll .001$	13.0
F0	DFT coeff. 9	$\ll .001$	9.1
F0	DFT coeff. 6	$\ll .001$	8.9
F0	Dist. max., arith. mean	$\ll .001$	6.8
F0	DFT coeff. 10	$\ll .001$	6.7
F0	IQR 1–3	$\ll .001$	6.4

(b) manual BP

Table 4: Relevance of functionals of pitch (F0) and F0 delta regression coefficients (Δ F0): p-values and t-statistics obtained on the evaluation database. IQR: inter-quartile ranges.

6. CONCLUSIONS

In singer-independent evaluation on a real-life database spanning different musical genres from pop to opera, we investigated different approaches to automatic recognition of vibrato in recorded polyphonic music from F0 contours and other low-level descriptors, using segment-wise functionals. It turned out that while the conventional approach of using the F0 discrete spectrum dramatically suffers from pitch estimation errors due to multiple present sources of accompaniment, percentiles of the delta regression (discrete derivative) of the pitch contour are highly robust.

Future work should focus on dynamic spotting of vibrato singing in recorded polyphonic music, integrating the proposed classification framework into a fully automatic system; this could be achieved by using note on-set detection, as a first stage, or dynamic modeling of segment-wise functionals.

7. REFERENCES

[1] T. L. Nwe and H. Li, "Exploring vibrato-motivated acoustic features for singer identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 519–530, 2007.

[2] C. Drioli and R. Di Federico, "Toward an integrated sound analysis and processing framework for expressiveness rendering," in *Proc. of ICMC*, Ann Arbor, Michigan, 1998, pp. 175–178.

[3] H.-S. Pang and D.-H. Yoon, "Automatic detection of vibrato in monophonic music," *Pattern Recognition*, vol. 38, no. 7, pp. 1135–1138, 2005.

[4] N. Amir, O. Michaeli, and O. Amir, "Acoustic and perceptual assessment of vibrato quality of singing students," *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 144–150, 2006.

[5] I. Luengo, I. Saratxaga, E. Navas, I. Hernáez, J. Sanchez, and I. Sainz, "Evaluation of Pitch Detection Algorithms Under Real Life Conditions," in *Proc. of ICASSP*, 2007, pp. 1057–1060.

[6] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[7] D. J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257–264, 1988.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, Florence, Italy, October 2010, pp. 1459–1462, ACM.

[9] S. J. Young, G. Evermann, M. J. F. Gales, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK book version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[10] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011—The First International Audio/Visual Emotion Challenge," in *Proc. First International Audio/Visual Emotion Challenge and Workshop (AVEC) held in conjunction with ACII*, LNCS, pp. 415–424. Springer, Memphis, TN, USA, 2011.

[11] N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," *Machine Learning*, pp. 161–205, May 2005.

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