Proceedings of the 2^{nd} AES Workshop on Intelligent Music Production, London, UK, 13 September 2016

AUTOMATIC BEST TAKE DETECTION FOR ELECTRIC GUITAR AND VOCAL STUDIO RECORDINGS

Carsten Bönsel, Jakob Abeßer, Sascha Grollmisch, Stylianos Ioannis Mimilakis

Fraunhofer Institute for Digital Media Technology (IDMT), Ilmenau, Germany

ABSTRACT

In the course of music recording sessions, the same vocal or instrumental passages are usually performed several times. However, only the best takes are chosen and further processed. Especially for lead vocals and solo instruments, the quantity of recorded material can be overwhelming, which makes the selection process time-consuming. Our goal is to automate and objectify this procedure in order to assist music producers for a faster decision making. The task of automatic best take detection is constrained to monophonic lines of electric guitar and singing voice in popular music. Assuming realistic scenarios during recording sessions, the proposed system requires only a synchronized click track and a backing track with accompanying instruments to be available for analysis.

1. INTRODUCTION

In the context of music studio recordings, our study deals with the question how the selection of the best take could be assisted by means of autonomous Music Information Retrieval (MIR) techniques. Recent publications [1–5] cover several relevant aspects of music performance analysis such as intonation and timing. However, the process of best take detection itself is—to the best knowledge of the authors—still an unexplored research field.

Our primary goal is to automatically produce a ranking of a given set of the recorded takes, ordered from best to worst. The ordering dimension which has to be estimated is denoted as music performance quality (MPQ), which according to Williamon and Valentine is defined as the overall presented and subject-specific ability. This ability consists of a defined collectivity of metrics, describing the three aspects of (1) musical understanding, (2) communicative ability, and (3) technical proficiency [6]. One of the main problems is that assessing MPQ is subjective to a certain extend. Therefore, one challenge is to identify objective criteria.

Another general problem in this area is the differentiation between musical shortcoming and intention. Similar shapes of phrasing are assessed differently in their MPQ. Depending on their underlying scheme or interpretive context, they can be perceived as both good or bad. For instance, unsystematic pitch modulation is usually perceived as tonal instability (and hence bad), while small, cyclic, regular pitch modulation—known as *vibrato*—is mostly interpreted as indication for a sophisticated technical proficiency. The same is true for time variation: If tempo is changed systematically, this behaviour is identified as *micro-timing* instead of rhythm inaccuracies. MPQ is highly context-sensitive in general. A heavy rough vocal timbre is appropriate in a rock song, but might be considered as misplaced in a pop ballad, for example.

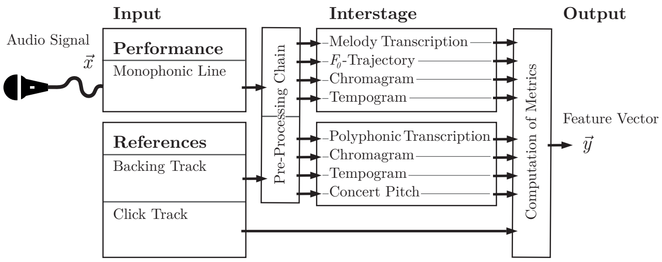

Figure 1: Structure & signal flow of the proposed system

2. PROPOSED FRAMEWORK

As shown in Figure 1, the proposed framework requires three types of input data—the audio from the musician's performance as well as the backing track and a click track providing harmonic and metrical information. Additionally, the equal tempered scale (ETS) is used as a reference. Automatic melody transcription and fundamental frequency (f_0) contour estimation are performed prior to the feature extraction [7]. Based on these melody representations, a set of timing and intonation features are computed.

The feature set was built up systematically based on a formed taxonomy to assure that all relevant areas are covered by at least one feature. Therefore, we followed the recommendation from [8] in designing distinct rubrics to form a rule set for the assessment task. In accordance with the four domains tonality/pitch (including melody and harmony), rhythm, intensity, and timbre, it is assumed that MPQ can be derived from the four musical rubrics of intonation, timing, dynamics, and sounding.

In a preliminary study, we conducted expert interviews. The experts' subject-specific knowledge provided further insight to the task at hand. Since the professional audio engineers confirmed our assumption that dynamics and sounding were highly subjective, only *intonation* and *timing* are further pursued. Both can be analyzed locally (note-wise),

Timing	**Intonation**
Local (note-wise)	
Note onset accuracy	Pitch accuracy
Note offset accuracy	Pitch stability
Note duration accuracy	Pitch drift amount
	Scooping amount
	Vibrato amount
Regional (pattern-wise)	
Relative timing	Relative intonation
Global (segment-wise)	
Overall static time shift	Overall static pitch shift

Table 1: Defined metrics for the systematic development of features

regionally (pattern-wise), and globally (segment-wise) which leads to the metrics shown in table 1.

Scooping means that singers slide into notes, starting each phrase on a low or indeterminate pitch beneath the note, then correcting it. While a scooped tone is well intonated, about 60 percent of the time the right pitch is not reached. *Pitch drift* is the opposed shape of a similar pitch modulation. While the correct pitch is reached at the beginning, the singer is drifting down in his intonation.

Based on the literature of related work, we chose four different intermediate musical representations. Such features include tempogram, distance, time quantization costs to binary / ternary grids pitch class histogram distances, pitch quantization costs towards the equal tempered scale, pitch stability measure, (f_0) slope and vibrato likelihood.

3. EVALUATION

The feature extraction step described in the previous section lead to a 88-dimensional feature vectors. In a first step, an ordinal classification is applied to determine the rank. For this purpose, several machine learning approaches have been attempted, including Support Vector Machine (SVM) using different kernel functions, Gaussian Mixture Models (GMM), and different types of regression. For our data set, partial least square regression (PLSR) performed best w.r.t. Kendall rank correlation (Kendall's τ). Using a novel dataset of 300 short monophonic guitar and vocal audio snippets with a total duration of 78 minutes, we achieve best performance values of $\tau = 0.68$ (guitar) and $\tau = 0.53$ (vocals).

4. CONCLUSION

The proposed system achieves good performance for the newly-defined task of best take detection. In general, the difficulty to differentiate between intention and deficiency remains the main challenge of the proposed task. Additionally, the amount of training data could be further increased to better represent different performance levels and music styles. The results for guitar are better than for vocals. One reason for this behavior is the higher error rate of the automatic transcription for vocals. While guitar transcriptions reach a F-measure of 0.91, for vocal transcription, the F-measure is merely 0.70. Transcription errors propagate to the feature extraction stage. Furthermore, we observed that even trained human raters does not agree in all cases.

5. REFERENCES

[1] Y. Lin, W.-C. Chang, and A. W. Su, "Quantitative evaluation of violin solo performance," in *Proceedings of the 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2013.

[2] P. Larrouy-Maestri, Y. Lévêque, D. Schön, A. Giovanni, and D. Morsomme, "The evaluation of singing voice accuracy: A comparison between subjective and objective methods," *Journal of Voice*, vol. 27, no. 2, pp. 259.e1–259.e5, 2013.

[3] P. Larrouy-Maestri and D. Morsomme, "Criteria and tools for objectively analysing the vocal accuracy of a popular song," *Logopedics Phoniatrics Vocology*, vol. 39, no. 1, pp. 11–18, 2014.

[4] 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.

[5] J. Abeßer, J. Hasselhorn, C. Dittmar, A. Lehmann, and S. Grollmisch, "Automatic quality assessment of vocal and instrumental performances of ninth-grade and tenth-grade pupils," in *Proceedings of the 10th International Symposium on Computer Music Modelling and Retrieval (CMMR)*, 2013.

[6] A. Williamon and E. Valentine, "Quantity and quality of musical practice as predictors of performance quality," *British Journal of Psychology*, vol. 91, pp. 353–376, 2000.

[7] K. Dressler, "Pitch estimation by the pair-wise evaluation of spectral peaks," in *Proceedings of the 42nd AES International Conference on Semantic Audio*, 2011.

[8] B. C. Wesolowski, "Understanding and developing rubrics for music performance assessment," *Music Educators Journal*, vol. 98, no. 3, pp. 36–42, 2012.