Melody Transcription from Music Audio: Approaches and Evaluation

Graham E. Poliner[1], Daniel P.W. Ellis[1], Andreas F. Ehmann[2], Emilia Gómez[3], Sebastian Streich[3], Beesuan Ong[3],

[1]LabROSA
Columbia University
New York, NY USA

[2] Deptartment of Electrical & Computer Engineering
University of Illinois at Urbana-Champaign
Urbana-Champaign, IL USA

[3]Music Technology Group
Universitat Pompeu Fabra,
Barcelona, Spain

Abstract— **Although the process of analyzing an audio recording of a music performance is complex and difficult even for a human listener, there are limited forms of information that may be tractably extracted and yet still enable interesting applications. We discuss melody – roughly, the part a listener might whistle or hum – as one such reduced descriptor of music audio, and consider how to define it, and what use it might be. We go on to describe the results of full-scale evaluations of melody transcription systems conducted in 2004 and 2005, including an overview of the systems submitted, details of how the evaluations were conducted, and a discussion of the results. For our definition of melody, current systems can achieve around 70% correct transcription at the frame level, including distinguishing between the presence or absence of the melody. Melodies transcribed at this level are readily recognizable, and show promise for practical applications.**

I. Introduction

Listeners respond to a wealth of information in music audio, and can be very sensitive to the fine details and nuances that can distinguish a great performance. Ever since the emergence of digital signal processing, researchers have been using computers to analyze musical recordings, but it has proven more challenging than expected to recognize the kinds of aspects, such as notes played and instruments present, that are usually trivial for listeners. Among these tasks, automatic transcription – converting a recording back in to the musical 'score', or list of note times and pitches, that the performer may have been reading – is a popular task: music students can perform transcription very effectively (after suitable training), but, despite a pretty clear understanding of the relationship between harmonics in the signal and perceived pitches, full transcription of multiple, overlapping instruments has proven elusive. Stretching back into the 1970s, a long thread of research has gradually improved transcription accuracy and reduced the scope of constraints required for success ([24], [21], [19], [28], [20], [22] among many others), but we are still far from a system that can automatically and accurately convert a recording back into a set of commands that would replicate it on a music synthesizer.

The basic problem is that while the pitch of a single musical note is consistently represented as a waveform with a more or less stable periodicity (giving rise to a set of harmonics at integer multiples of a fundamental under Fourier analysis), ensemble music will frequently include episodes where four or more notes are overlapping in time, and moreover the fundamentals of these notes may be in simple integer ratios, meaning their harmonics actually coincide, giving complex patterns of constructive and destructive interference in a narrowband spectral analysis; this harmonic mingling appears to be at the core of musical harmony. In view of this difficulty, researchers have considered alternative formulations that might be more practical than full transcription while still supporting some of the applications that transcription would enable. Goto suggested identifying just a single, dominant periodicity over the main spectral range of music (plus one more in the low frequencies, corresponding to the bass line), which he referred to as "Predominant-F0 Estimation" or PreFEst [16], [13]. This restriction allowed both a tractable implementation (running in real-time even in 1999) and a musically-interesting description that gave recognizable 'sketches' of many popular music examples.

Although Goto was careful not to blur the distinction, in most cases his predominant pitch was recognizable as the melody of the music, and this paper is concerned specifically with the problem of extracting the melody from music audio. Providing a strict definition of the melody is, however, no simple task: it is a musicological concept based on the judgment of human listeners, and will not, in general, be uniquely defined for all recordings. Roughly speaking, the melody is the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognize as being the 'essence' of that music when heard in comparison. In many cases, listeners find it easy to identify the melody; in particular, much of popular music has a 'lead vocal' line, a singer whose voice is the most prominent source in the mixture, and who is singing the melody line. But even in classical orchestral music, or richly polyphonic piano compositions, in very many cases a single, prominent melody line can be agreed upon by most listeners. Thus, while we are in the dangerous position of setting out to quantify the performance of automatic systems seeking to extract something that is not strictly defined, there is some hope we can conduct a meaningful evaluation.

Figure 1 gives an example of what we mean by a melody, and illustrates some of the difficulties of the problem of melody transcription. As discussed in section III, we have obtained a small number of recordings where the vocal line is presented alone (from the original multitrack recordings made in the studio). We assume this lead vocal constitutes the

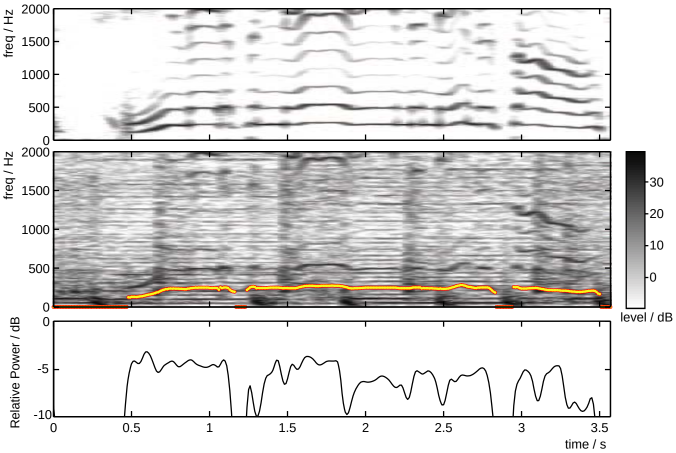

Fig. 1. Illustration of melody in polyphonic music. Top pane: narrowband spectrogram of vocal line (i.e. melody) from original multi-track recording. Middle pane: corresponding spectrogram of the full polyphonic recording, when all accompaniment has been mixed in. Bottom pane: Power of melody relative to full mix.

melody; its spectrogram (using a 100ms window in order to emphasize the harmonic structure of the lead voice) is shown in the top pane. The spectrogram below, however, is the full polyphonic recording with all the accompaniment instruments present. Clearly, the melody line is much less prominent, as confirmed by the bottom pane which shows the power of the melody signal compared to the full mix, as a function of time.

Accurate melody transcription would make possible numerous applications: One obvious direction arises from the popular paradigm of "Query-by-humming" [10], [2] which aims to help users find a particular piece of music based on a hummed or sung extract. By our definition, we can assume that the queries will be fragments of melody, but if the database consists of full, polyphonic recordings we cannot expect the query to resemble the recording in any broad sense. Melody transcription would allow us to describe each database item in terms of its melody, and match queries in that domain. In fact, for this application, melody transcription may be preferable to full, polyphonic transcription, since it also provides a necessary solution to the problem of identifying the melody line within the full set of notes being played.

Other applications for melody transcription include anywhere that a reduced, simplified representation of music might be advantageous, such as clustering different variations of the same piece, or analyzing common musicological primitives. Melodies can also be a kind of thumbnail or cartoon of a full recording e.g. for limited-capacity devices such as some cellphones. Score following, where a complex recording is temporally aligned to a known performance score, might also be easier and more successful in such a reduced, but still informative, domain.

The remainder of this paper is organized as follows: In the next section, we present an overview of the different approaches taken to melody transcription, based on the submissions made to the two annual evaluations of this task we have conducted. Section III then gives details of these evaluations, describing both how the materials were prepared, and what metrics we used. Then, in section IV we present the results of the evaluations, and, as far as possible, make observations concerning the performance of the different approaches. We mention future directions and draw conclusions in section V.

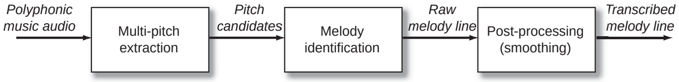

Fig. 2. Basic processing structure underlying all melody transcription systems.

II. Approaches to Melody Transcription

Melody transcription is strongly related to pitch tracking, which itself has a long and continuing history (for reviews, see [18], [17], [3]). In the context of identifying melody within multi-instrument music, the pitch tracking problem is further complicated because although multiple pitches may be present at the same time, at most just one of them will be the melody. Thus, all approaches to melody transcription face two problems: identifying a set of candidate pitches that appear to be present at a given time, then deciding which (if any) of those pitches belongs to the melody. Note that the task of detecting whether the melody is active or silent at each time, although seemingly secondary, turned out to be a major factor in differentiating performance in the evaluations. Finally, a sequence of melody estimates can be post-processed, typically to remove spurious notes or otherwise increase smoothness. Figure 2 shows the basic processing sequence that more or less covers all the algorithms we will discuss.

The audio melody transcription competitions conducted in 2004 and 2005 (described in section III) attracted a total of 14 submissions – four in 2004 and ten in 2005. Of the algorithms evaluated in 2004, all but one were also represented in 2005, the exception being the autocorrelation-based scheme of Bello. Of the ten submissions in 2005, two were 'contrast' variants of other submissions, and one never delivered interpretable results due to system issues, leaving seven main algorithms to compare. These are listed in table I, which attempts to break down the description of the algorithms into several key dimensions. Systems are referred to by their first authors only, for brevity. The ordering of the algorithms in the table aims merely to highlight their similarities.

The first column, "Front end", concerns the initial signal processing applied to input audio to reveal the pitch content. The most popular technique is to take the magnitude of the short-time Fourier transform (STFT) – the Fourier transform of successive, windowed, snippets of the original waveform – denoted $|STFT|$ in the table, and commonly visualized as the spectrogram. Pitched notes appear as a 'ladder' of more-or-less stable harmonics on the spectrogram, a clear visual representation that suggests the possibility of automatic detection. Unlike the time waveform itself, $|STFT|$ is invariant to relative or absolute time or phase shifts in the harmonics because the STFT phase is discarded. This is convenient since perceived pitch has essentially no dependence on the relative phase of (resolved) harmonics, and it makes the estimation invariant to alignment of the analysis time frames. Since the

TABLE I

PRINCIPAL MELODY TRANSCRIPTION ALGORITHMS. SEE TEXT FOR DETAILS.

System	Front end	Multi-pitch	# pitches	Onset events	Post-processing	Voicing
Dressler [6]	$\lvert STFT\rvert$ + sines	Harmonic model fit	5	Fragments	Streaming rules	Melody + local thresh.
Marolt [23]	$\lvert STFT\rvert$ + sines	EM fit of tone models	> 2	Fragments	Proximity rules	Melody grouping
Goto [14]	Hier. $\lvert STFT\rvert$ + sines	EM fit of tone models	> 2	–	Tracking agents	(continuous)
Ryynänen [29]	Auditory + $\lvert STFT\rvert$	Harmonic Sieve	2	Note onsets	HMM	Background model
Poliner [27]	$\lvert STFT\rvert$	SVM Classifier	1	–	–	Global threshold
Paiva [26]	Auditory correlogram	Summary autocorrelation	> 2	Pitches	Pruning rules	Melody grouping
Vincent [35]	YIN / Time windows	Gen. model inference	5 / 1	–	HMM	(continuous)

frequency resolution of the STFT improves with temporal window length, these systems tend to use long windows, from 46 ms for Dressler, to 128 ms for Poliner. Goto uses a hierarchy of STFTs to achieve a multiresolution Fourier analysis, downsampling his original 16 kHz audio through 4 factor-of-2 stages to have a 512 ms window at his lowest 1 kHz sampling rate. Since musical semitones are logarithmically spaced with a ratio between adjacent fundamental frequencies of $2^{1/12} \approx 1.06$, to preserve semitone resolution down to the lower extent of the pitch range (below 100 Hz) requires these longer windows. Ryynänen uses an auditory model front-end to enhance and balance information across the spectrum, but then calculates the $|STFT|$ for each subband and combines them. Dressler, Marolt, and Goto further reduce their magnitude spectra by recording only the sinusoidal frequencies estimated as relating to prominent peaks in the spectrum, using a variety of techniques (such as instantaneous frequency [9]) to exceed the resolution of the STFT bins.

Two systems do not use the STFT: Paiva uses the Lyon-Slaney auditory model up to the summary autocorrelation [32], and Vincent uses a modified version of the YIN pitch tracker [4] to generate candidates for his later time-domain model inference. Both these approaches use autocorrelation, which also achieves phase invariance (being simply the inverse Fourier transform of $|STFT|^2$) but also has the attractive property of summing all harmonics relating to a common period into a peak at that period. The Lyon-Slaney system actually calculates autocorrelation on an approximation of the auditory nerve excitation, which separates the original signal into multiple frequency bands, then sums their normalized results; Paiva's multi-pitch detection involves simply choosing the largest peaks from this summary autocorrelation. Although YIN incorporates autocorrelation across the full frequency band, Vincent calculates this from the STFT representation, and reports gains from some degree of across-spectrum energy normalization. Interestingly, because the resolution of autocorrelation depends on the sampling rate and not the window length, Paiva uses a significantly shorter window of 20 ms, and considers periods only out to 9 ms lag (110 Hz).

The next column, "Multi-pitch", addresses how the systems deal with distinguishing the multiple periodicities present in the polyphonic audio, and the following column, "# pitches", attempts to quantify how many simultaneous pitches can be reported at any time. For systems based on $|STFT|$, the problem is to identify the sets of harmonics and properly credit the energy or salience of each harmonic down to the appropriate fundamental – even though there need not be any energy at that fundamental for humans to perceive the pitch. This generally reduces to a 'harmonic sieve' [11], [8], which, in principal at least, considers every possible fundamental and integrates evidence from every predicted harmonic location. One weakness with this approach is its susceptibility to reporting a fundamental one octave too high, since if all the harmonics of a fundamental frequency f_0 are present, then the harmonics of a putative fundamental $2f_0$ will also be present. Ryynänen implements a harmonic sieve more or less directly, but identifies lower fundamentals first, then modifies the spectrum to remove the energy associated with the low pitch, thereby removing evidence for octave errors. Goto proposed a technique for estimating weights over all possible fundamentals to jointly explain the observed spectrum, which effectively lets different fundamentals compete for harmonics, based on Expectation-Maximization (EM) re-estimation of the set of unknown harmonic-model weights; this is largely successful in resolving octave ambiguities [14]. Marolt modifies this procedure slightly to consider only fundamentals that are equal to, or one octave below, actual observed frequencies, and then integrates nearby harmonics according to perceptual principles. The results of these (and an apparently similar procedure in Dressler) are weights assigned to every possible pitch, most of which are very small; the few largest values are taken as the potential pitches at each frame, with typically 2 to 5 simultaneous pitches being considered.

Poliner takes a radical approach of feeding the entire Fourier transform magnitude at each time slice, after some local normalization, into a Support Vector Machine (SVM) classifier. This classifier has previously been trained on many thousands of example spectral slices for which the appropriate melody note is known (e.g. through manual or human-corrected transcription of the original audio), and thus it can be assumed to have learned both the way in which pitches appear as sets of harmonics, and also how melody is distinguished from accompaniment, to the extent that this is evident within a single short-time window. This approach willfully ignores prior knowledge about the nature of pitched sounds, on the principle that it is better to let the machine learning algorithm figure this out for itself, where possible. The classifier is trained to report only one pitch – the appropriate melody – for each frame, quantized onto a semitone scale, and this was used, without further processing, as the pitch estimate in the evaluated system.

Although Vincent starts with an autocorrelation to get up to 5 candidate periods for consideration, the core of his system is a generative model for the actual time-domain waveform within each window that includes parameters for fundamental frequency, overall gain, amplitude envelope of the harmonics, the phase of each harmonic, and a background noise term that scales according to local energy in a psychoacoustically-derived manner. The optimal parameters are inferred for each candidate fundamental, and the one with the largest posterior probability under the model is chosen as the melody pitch at that frame.

The next column, "Onset events", reflects that only some of the systems incorporate sets of distinct objects – individual notes or short strings of notes – each with a distinct start and end time, internal to their processing. Three systems, Goto, Poliner, and Vincent, simply decide a single best melody pitch at every frame and do not attempt to form them into higher note-type structures. Dressler and Marolt, however, take sets of harmonics similar to those in Goto's system, but track the amplitude variation to form distinct fragments of more-or-less continuous pitch and energy that are then the basic elements used in later processing (since there may still be multiple elements active at any given time). Paiva goes further to carefully resolve his continuous pitch tracks into piecewise-constant frequency contours, thereby removing effects such as vibrato (pitch modulation) and slides between notes to get something closer to the underlying, discrete melody sequence (the evaluation, however, was against ground truth giving the actual fundamental rather than the intended note, so Paiva's system eventually reported this earlier value).

Ryynänen uses a hidden Markov model (HMM) providing distributions over features including an 'onset strength' related to the local temporal derivative of total energy associated with a pitch. The first, "attack", state models the sharp jump in onset characteristics expected for new notes, although a bimodal distribution also allows for notes that begin more smoothly; the following "sustain" state is able to capture the greater salience (energy), narrower frequency spread, and lesser onset strength associated with continuing notes. Thus, new note events can be detected simply by noting transitions through the onset state for a particular note model in the best-path (Viterbi) decoding of the HMM.

The second-to-last column, "Post-processing", looks at how raw (multi) pitch tracks are further cleaned up to give the final melody estimates. In the systems of Dressler, Marolt, and Paiva, this involves choosing a subset of the note or note fragment elements to form a single melody line, including gaps where no melody note is selected. In each case, this is achieved by sets of rules that attempt to capture the continuity of good melodies in terms of energy and pitch (i.e. avoiding or deleting large, brief, frequency jumps). Rules may also include some musical insights, such as preference for a particular pitch range, and for the highest or lowest (outer) voices in a set of simultaneous pitches (a polyphony). Although Goto does not have an intermediate stage of note elements, he does have multiple pitch candidates to choose between, which he achieves via a set of interacting "tracking agents" – alternate hypotheses of the current and past pitch – which compete to acquire the new pitch estimates from the current frame, and live or die based on a continuously-updated penalty that reflects the total strength of the past pitches they represent; the strongest agent determines the final pitch reported.

Ryynänen and Vincent both use HMMs to limit the dynamics of their pitch estimates i.e. to provide a degree of smoothing that favors slowly-changing pitches. Ryynänen simply connects his per-note HMMs through a third, noise/background, state, and also has the opportunity to include musicologically-informed transition probabilities that vary depending on an estimate of the current chord or key [34]. Vincent uses an HMM simply to smooth pitch sequences, training the transition probabilities as a function of interval size from the ground-truth melodies in the 2004 evaluation set.

The final column, "Voicing", considers how, specifically, the systems distinguish between intervals where the melody is present and those where it is silent (gaps between melodies). Goto and Vincent simply report their best pitch estimate at every frame and do not admit gaps. Poliner's basic pitch extraction engine is also continuous, but this is then gated by a separate melody detector; a simple global energy threshold over an appropriate frequency range was reported to work as well as a more complex scheme based on a trained classifier. As discussed above, the selection of notes or fragments in Dressler, Marolt, and Paiva naturally leads to gaps where no suitable element is selected; Dressler augments this with a local threshold to discount low-energy notes.

III. The Melody Evaluations

As described above, there are many approaches to the melody transcription problem. Until recently though, a number of obstacles such as the lack of a standarized test set or consensus regarding evaluation metrics impeded an objective comparison of these systems. In 2004, the Music Technology Group at the University of Pompeu Fabra proposed and hosted a number of audio description contests in conjunction with the International Conference on Music Information Retrieval (ISMIR). These evaluations which included contests for melody transcription, genre classification/artist identification, tempo induction, and rhythm classification evolved into the Music Information Retrieval Evaluation Exchange (MIREX) [5] which took place during the summer of 2005, organized and run by Columbia University and the University of Illinois at Urbana-Champaign. In this section, we examine the steps that have been taken toward an objective comparison of melody transcription systems.

A. Evaluation Material

Although a great deal of music is available in a digital format, the number of corresponding transcriptions time-aligned to the audio is rather limited. Recently, Goto et al. prepared the Real World Computing (RWC) Music Database [15] which contains 315 recordings of musical pieces along with accompanying standard MIDI files – descriptions of the note events rounded to the nearest semitone. Although the RWC database has proven to be a very valuable resource, discretizing audio to the nearest semitone omits a significant amount of the expressive detail (e.g. vibrato and glide transitions) that is critical to musicological analysis. In addition, the problem of identifying the predominant melody given a complete transcription is still an open research problem [25], [30]. As such, novel sets of recording-transcription pairs were required in order to perform real-world melody transcription evaluations.

Trained musicians are capable of generating detailed transcriptions from recorded audio; however, the process is often difficult and time consuming for ensemble pieces. As an alternative to labeling the audio by hand, standard recording conventions may be exploited in order to facilitate the creation of reference transcriptions. In many cases, music recordings are made by layering a number of independently recorded audio tracks. In some instances, artists (or their record companies) distribute the full set of multi-track recordings, or a reduced set (e.g. separate vocal and instrumental tracks), as part of a single release. The monophonic lead voice recordings can be used to create ground truth for the melody in the full ensemble music, since the solo voice can usually be tracked with high accuracy by standard pitch tracking systems [33], [1], [4]. In both evaluations the test sets were supplemented with synthesized audio (e.g. MIDI); however, the contest organizers sought to limit the inclusion of these recordings wherever possible since the the reduced acoustic complexity may lead to poor generalization on commercial recordings.

A description of the data used in the 2004 evaluation is displayed in table II. The test set is made up of 20 monaural audio segments (44.1 kHz sampling rate, 16 bit PCM) across a diverse set of musical styles. The corresponding reference data was created by using SMSTools [1] to estimate the fundamental frequency of the isolated, monophonic, melody track at 5.8 ms steps. As a convention, the frames in which the main melody is unvoiced are labeled 0 Hz. The transcriptions were manually verified and corrected in order to ensure the quality of the reference transcriptions. Prior to the evaluation, half of the test set was released for algorithm development, and the remainder was released shortly after the competition.

TABLE II

SUMMARY OF THE TEST DATA USED IN THE 2004 MELODY EVALUATION. EACH CATEGORY CONSISTS OF 4 EXCERPTS, EACH ROUGHLY 20 S IN DURATION. THE 8 SEGMENTS IN THE *Daisy* AND *MIDI* CATEGORIES WERE GENERATED USING A SYNTHESIZED LEAD MELODY VOICE, AND THE REMAINING CATEGORIES WERE GENERATED USING MULTI-TRACK RECORDINGS.

Category	**Style**	**Melody Instrument**
Daisy	Pop	Synthesized voice
Jazz	Jazz	Saxophone
MIDI	Folk (2), Pop (2)	MIDI instruments
Opera	Classical opera	Male voice (2), Female voice (2)
Pop	Pop	Male Voice

Since the 2004 data was distributed after the competition, an entirely new test set of 25 excerpts was collected for the 2005 evaluation. The same audio format was used as in the 2004 evaluation; however, the ground-truth melody transcriptions were generated at 10 ms steps using the ESPS get_f0 method implemented in WaveSurfer [31]. The fundamental frequency estimates were manually verified and corrected using the graphical user interface as displayed in figure 3. Prior to the contest, a representative set of 3 segments was provided for algorithm tuning; however, the 25 test songs have been reserved for future evaluations and, therefore, have not been publicly distributed.

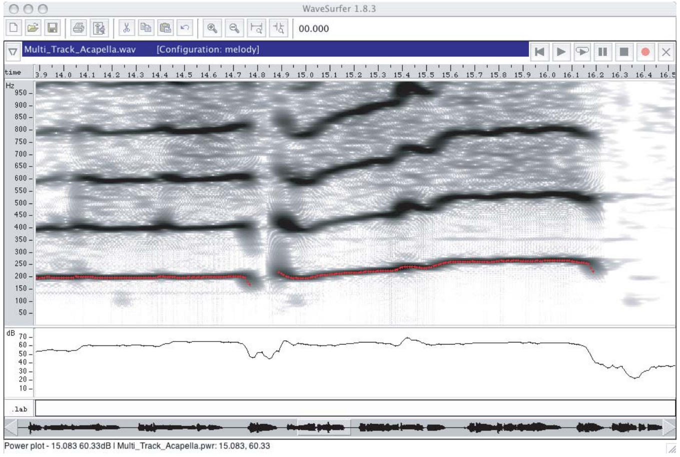
WaveSurfer 1.8.3
00.000
Multi_Track_Acapella.wav [Configuration: melody]
time
Hz
950
900
850
800
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50
dB
70
60
50
40
30
20
10
.lab
Power plot - 15.083 60.33dB | Multi_Track_Acapella.pwr: 15.083, 60.33

Fig. 3. Screenshot of the semi-automatic melody annotation process. Minor corrections were made to the output of a monophonic pitch tracker on the isolated melody track, and the reference transcriptions were time-aligned to the full ensemble recording by identifying the maximum cross-correlation between the melody track and the ensemble.

TABLE III

SUMMARY OF THE TEST DATA USED IN THE 2005 MELODY EVALUATION.

Melody Instrument	**Style**
Human voice (8 f, 8 m)	R&B (6), Rock (5), Dance/Pop (4), Jazz (1)
Saxophone (3)	Jazz
Guitar (3)	Rock guitar solo
Synthesized Piano (3)	Classical

As displayed in table III, the 2005 test data was more heavily biased toward a pop-based corpora rather than uniformly weighting the segments across a number of styles/genres as in the 2004 evaluation. The shift in the distribution was motivated both by the relevance of commercial applications for music organization, and by the availability of multi-track recordings in the specified genres. Since the 2005 test set is more representative of real-world recordings, it is inherently more complex than the preceeding test set.

B. Evaluation Metrics

Algorithms submitted to the contests were required to estimate the fundamental frequency of the predominant melody on a regular time grid. An attempt was made to evaluate the lead voice transcription at the lowest level of abstraction, and as such, the concept of segmenting the fundamental frequency predictions into notes has been largely omitted from consideration. The metrics used in each of the evaluations were agreed upon by the participants in a discussion period prior to algorithm submission. In this subsection, we present an evolutionary description of the evaluation metrics.

1) 2004 Evaluation Metrics: For the 2004 evaluation, the submitted algorithms output a single prediction combining fundamental frequency estimation and voicing detection at each instant. The submissions were evaluated against two metrics: raw transcription concordance, and chroma transcription concordance[1]. A final ranking of the submitted algorithms was determined by averaging the scores of the fundamental frequency and chroma transcription concordance.

The raw transcription concordance is a frame-based comparison of the estimated fundamental frequency to the reference fundamental frequency on a logarithmic scale. Both the estimated and reference fundamental frequency are converted to the cent scale:

$$f_{cent} = 1200 \cdot \left(log_2 \left(\frac{f_{Hz}}{13.75\ Hz} \right) - 0.25 \right), \qquad (1)$$

in order to compare the estimated fundamental to the reference pitch on a logarithmic scale, and the frame concordance error in frame n is measured by the absolute difference between the estimated and reference pitch value:

$$err_n = \begin{cases} 100 & \text{for } |f_{cent}^{est}[n] - f_{cent}^{ref}[n]| \geq 100 \\ |f_{cent}^{est}[n] - f_{cent}^{ref}[n]| & \text{otherwise} \end{cases} . \qquad (2)$$

Thus, the overall transcription concordance for a specific segment is given by the average concordance over all frames:

$$score = 100 - \frac{1}{N} \sum_{n=1}^{N} err_n. \qquad (3)$$

Unvoiced frames are included in the overall concordance score by binary assignment.

Octave transpositions and other errors in which the estimated pitch is off by an integer (sub)multiple of the reference pitch, are generally common in fundamental frequency estimation. As such, the chroma transcription concordance forgives octave errors by folding both the estimated and reference pitch into a single octave of 12 semitones before calculating the absolute difference score as above.

2) 2005 Evaluation Metrics: The structure of the melody competition was updated in 2005 to enable participants to perform pitch estimation and voicing detection independently i.e. each algorithm could give its best guess for a melody pitch even for frames that it reported as unvoiced. This modification to the evaluation allowed for more detailed insight into the structure of each system and encouraged participation by systems that do not consider melody voicing detection. In addition, the scoring metric for the voiced frames was relaxed to account for the precision limits in generating the reference transcription. A brief description of the updated evaluation metrics is provided below:

- The algorithms were ranked according to the **overall transcription accuracy**, a measure that combines the pitch transcription and voicing detection tasks. It is defined as the proportion of frames correctly labeled with both raw pitch accuracy and voicing detection.
- The **raw pitch accuracy** is defined as the proportion of voiced frames in which the estimated fundamental frequency is within ±1/4 tone of the reference pitch (including the pitch estimates for frames detected as unvoiced). Whereas the 2004 metric penalized slight deviations from the reference frequency, the updated pitch accuracy metric grants equal credit to all estimations within a quarter tone of the reference frequency in order to account for small frequency variations in the reference transcriptions.
- The **raw chroma accuracy** is defined in the same manner as the raw pitch accuracy; however, both the estimated and reference frequencies are mapped into a single octave in order to forgive octave transpositions.
- The **voicing detection rate** is the proportion of frames labeled voiced in the reference transcription that are estimated to be voiced by the algorithm.
- The **voicing false alarm rate** is the proportion of frames that are not voiced (melody silent) according to the reference transcription that are estimated to be voiced by the algorithm.
- The **discriminability** d' is a measure of a detector's sensitivity that attempts to factor out the overall bias toward labeling any frame as voiced (which can move both detection and false alarm rates up and down in tandem). Any combination of detection rate and false alarm rate can arise from setting a particular threshold on a scalar decision variable generated by two overlapping unit-variance Gaussians; d' is the separation between the means of those Gaussians required to achieve the given detection rates. A larger value indicates a detection scheme with better discrimination between the two classes [7].

The performance of each algorithm was evaluated on the 25 test songs, and the results of the evaluation are presented in the next section.

IV. Results and Discussion

In order to reflect the most recent research, we present only the melody transcription results from the 2005 evaluation; detailed results from the 2004 evaluation are available in [12].

The results of the melody transcription evaluation are provided in table IV. Looking first at the overall accuracy metric, we note that system proposed by Dressler outperformed the other submissions by a significant margin. As displayed in the top pane of figure 7, the Dressler system was the best algorithm on 17 of the 25 test songs and performed consistently across all musical styles. Figure 7 also illustrates how inconsistent transcription accuracy significantly affected the overall scoring for a few of the participants, most notably Marolt and Vincent. We summarize the relevant statistics pertaining to the overall accuracy of each system in figure 4. Recall that the submissions made by Goto and Vincent did not include voicing detection and as such cannot be directly compared to the other systems on overall accuracy.

If instead we examine the transcription stages independently, the results of the evaluation are more equivocal. With respect to the raw pitch accuracy, three systems performed within a statistically insignificant margin, and all of the submissions performed within 10% of each other. Considering

[1] An additional metric evaluating note-level melodic similarity was proposed; however, the results of the evaluation are not discussed in this paper owing to a lack of participation.

TABLE IV

RESULTS OF THE FORMAL MIREX 2005 AUDIO MELODY EXTRACTION EVALUATION. SUBMISSIONS MARKED WITH A * ARE NOT DIRECTLY COMPARABLE TO THE OTHERS FOR THE VOICING METRICS AND OVERALL ACCURACY BECAUSE THOSE SYSTEMS DID NOT PERFORM VOICED/UNVOICED DETECTION.

Rank	Participant	Overall Accuracy	Raw Pitch	Raw Chroma	Voicing Detection	Voicing FA	Voicing d'	Runtime / s
1	Dressler	**71.4%**	68.1%	71.4%	81.8%	17.3%	**1.85**	32
2	Ryynänen	64.3%	**68.6%**	**74.1%**	90.3%	39.5%	1.56	10970
3	Poliner	61.1%	67.3%	73.4%	91.6%	42.7%	1.56	5471
3	Paiva 2	61.1%	58.5%	62.0%	68.8%	23.2%	1.22	45618
5	Marolt	59.5%	60.1%	67.1%	72.7%	32.4%	1.06	12461
6	Paiva 1	57.8%	62.7%	66.7%	83.4%	55.8%	0.83	44312
7	Goto *	49.9%	65.8%	71.8%	99.9%	99.9%	0.59	211
8	Vincent 1 *	47.9%	59.8%	67.6%	96.1%	93.7%	0.23	62760
9	Vincent 2 *	46.4%	59.6%	71.1%	99.6%	96.4%	0.86	251

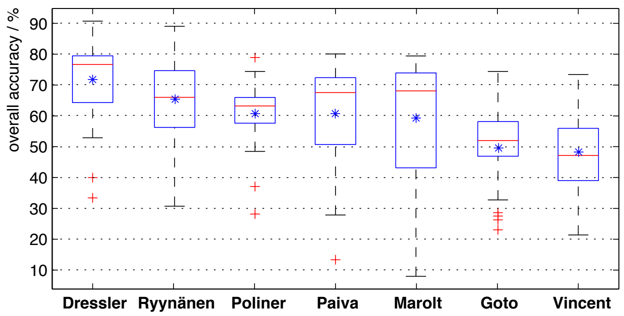

Fig. 4. Statistical summary of the overall accuracy results. The horizontal lines of the boxes denote the interquartile range and median. The star indicates the mean. The whiskers show the extent of the data, and outliers are indicated by + symbols.

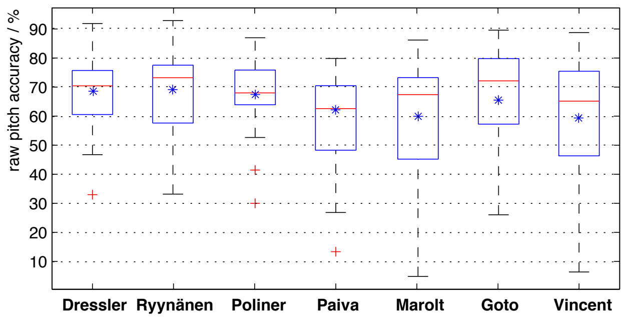

Fig. 5. Statistical summary of the raw pitch accuracy results. Symbols as for figure 4.

pitch estimation alone, Ryynänen's system was the best on average, and the Goto submission was the top performing algorithm on 12 of the 16 songs for which the lead melody instrument is the human voice. The raw pitch accuracy results for each song are displayed in the bottom pane of figure 7, and the summary statistics for the submissions are displayed in figure 5. In her MIREX submission, Dressler did not estimate a fundamental frequency for frames she labeled unvoiced, and as such, we can not make a direct comparison between her submission and the other systems on raw pitch accuracy. However, shortly after the results of the competition were released, Dressler submitted a modified algorithm that output fundamental frequency predictions for the unvoiced frames which resulted in a 1% improvement in raw pitch transcription accuracy over the value in table IV.

The raw chroma metric indirectly evaluates the candidate note identification stage and hints at the potential for improvment in post-processing. We note that the systems with $|STFT|$ based front end generally resulted in a higher raw chroma average, and that the rule-based post-processing implementations such as Dressler and Paiva minimized the difference between raw pitch and chroma accuracy. At this point, it seems as though the machine learning post-processing approaches do not sufficiently model the melody note transitions. In general, we expect the transitions to be limited to local steps; therefore, large jumps with short duration may be indicative of erroneous octave transpositions that could be filtered by post-processing.

Figure 6 displays note error histograms for a few of the submissions on training song #2, the song "Frozen" by Madonna. We observe that many of the errors are due to octave transpositions and harmonically related notes; however, these errors tend to be system specific. For instance, fundamental frequency tracking $|STFT|$ systems such as Dressler's submission tend to incorrectly estimate melody frequencies at twice the reference frequency – that is the reference note number plus 12 semitones. Although Ryynänen uses a similar front end system to Dressler, the errors generated by the musically constrained HMM were distributed over a two octave range and were often harmonically related notes on the circle of fifths. These systems contrast with the classification approach which exhibits a significant number of adjacent note errors due to discretizing estimates to the nearest semitone.

Upon examining example transcriptions, the stylistic differences between the different approaches become very pronounced. In Figure 8, we provide representative transcriptions from a few of the algorithms on the Madonna training file. Again we see that algorithms that track the fundamental frequency of the lead melody voice (e.g. Dressler) follow the reference transcription quite closely and provide a clear representation of the acoustic effects, whereas note modeling post-processing approaches and the classification-based system (Poliner) that discretize each estimate to the nearest semitone provide a representation that is more closely associated with

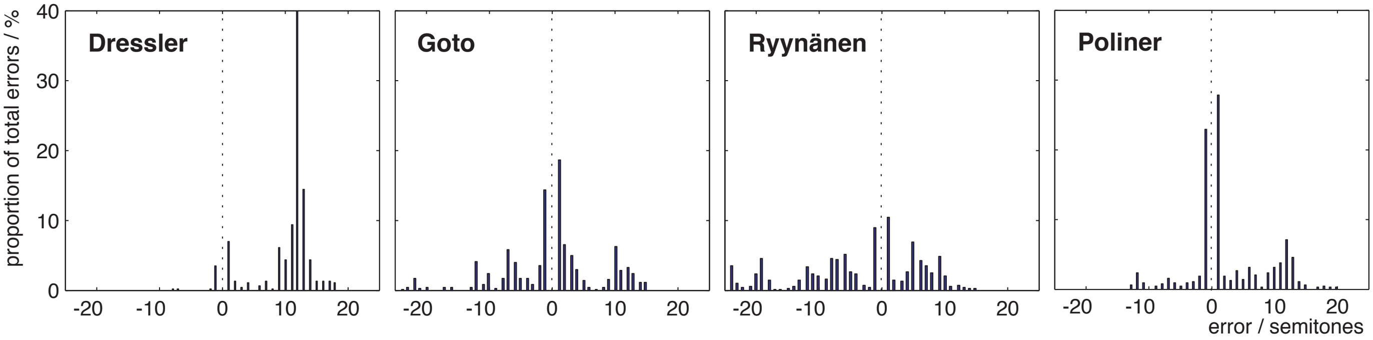

Fig. 6. Transcription error histograms where the relative frequency of errors is plotted against the number of semitones deviation from the reference pitch.

the note level of abstraction.

We can also identify a number of trends by looking at the raw pitch accuracy results across the different melody instruments. In general, the algorithms perform more similarly across the sung excerpts with average standard deviations of 7 and 6 on the female and male recordings respectively. In contrast, there is a large variance across the instrumental excerpts which highlights both the contextual difficulty of identifying the lead melody voice within an ensemble of similar instruments and the apparent overtraining bias toward sung melodies. The transcription results are consistently higher on songs for which the melody is well structured with a high foreground-to-background energy ratio such as the jazz excerpts, and many of the algorithms performed poorly on excerpts in which the lead voice is performed on a non-traditional melody instrument such as the guitar solos. The low piano transcription averages seem to support the notion that timbral variation may provide additional insight into the lead melodic voice.

While the voicing detection stage was somewhat of an afterthought for a number of the submissions (when it was considered at all), it proved to be the deciding feature in the evaluation. Dressler's approach of grouping melody phrases combined with a local energy threshold significantly outperformed the systems which considered either of the two independently. Using a fixed energy threshold alone generates false alarms when the melody is a smaller fraction of the total signal and false negatives when the melody is a larger fraction of the total signal. Conversely, the schemes that implemented melody grouping alone underestimated the total percentage of voiced frames in the evaluation. The key advantage in combining the melody grouping and threshold features appears to be a detection threshold that is invariant to the proportion of voiced melody frames. We note that the voicing detection and false alarm rate deviate slightly from 100% for the algorithms that did not consider voicing detection due to duration scaling artifacts.

Although it was not proposed as an evaluation metric, algorithm run-time is often of critical practical importance. The majority of the front end stages systems are quite similar in terms of complexity; however, the candidate pitch identification and post-processing stages vary significantly in terms of computational cost. The submitted algorithms differed in implementation from compiled code to functions in MATLAB. Although many of the submissions have not been optimized for efficiency, we see an enormous variation of over 1000:1 between the fastest and slowest systems – with the top ranked system also the fastest at under 0.1 times real time. This result underscores the feasability of using melody transcription as a tool for analyzing large music databases.

The results of the evaluation may also be used to gain insight into the quality of the test set. We expect, in general, a high degree of correlation between intersong performance as an indication of the discriminability of a given test song. For example, the first of the three Saxophone test samples provides a high degree of discriminability which is consistent with the overall results, while the third of the Guitar samples appears to provide a low degree of discriminability and is largely uncorrelated with the overall results of the evaluation – potentially an indication that the melody is ambiguous in the given context. We might hope to improve the quality of the test set for future evaluations by including additional songs across a more diverse set of genres.

V. Conclusions

The evaluations conducted as part of the 2004 and 2005 ISMIR conferences allowed a wide range of labs that had been independently studying melody transcriptions to come together and make a quantitative comparison of their approaches. As we have outlined, there were some significant algorithmic variations between the submissions, in terms of front-end, multipitch identification strategy, and post-processing. However, by factoring out the differences arising from the inclusion or omission of voicing detection, the raw pitch accuracy results show a surprisingly consistent performance, with all systems scoring between 60 and 70%. This perhaps suggests a distribution in the test set between 60% of frames which are quite easy, some intermediate difficulty, and a core of 30% of frames which are much harder, leading to a possible plateau in performance at this level.

At a more abstract level, the benefits of common, standardized evaluation are clearly shown by this effort and analysis. We aim to repeat the evaluation in 2006, and we are working to enhance the test set, metrics, and diagnostic analysis in light of our experiences to date.

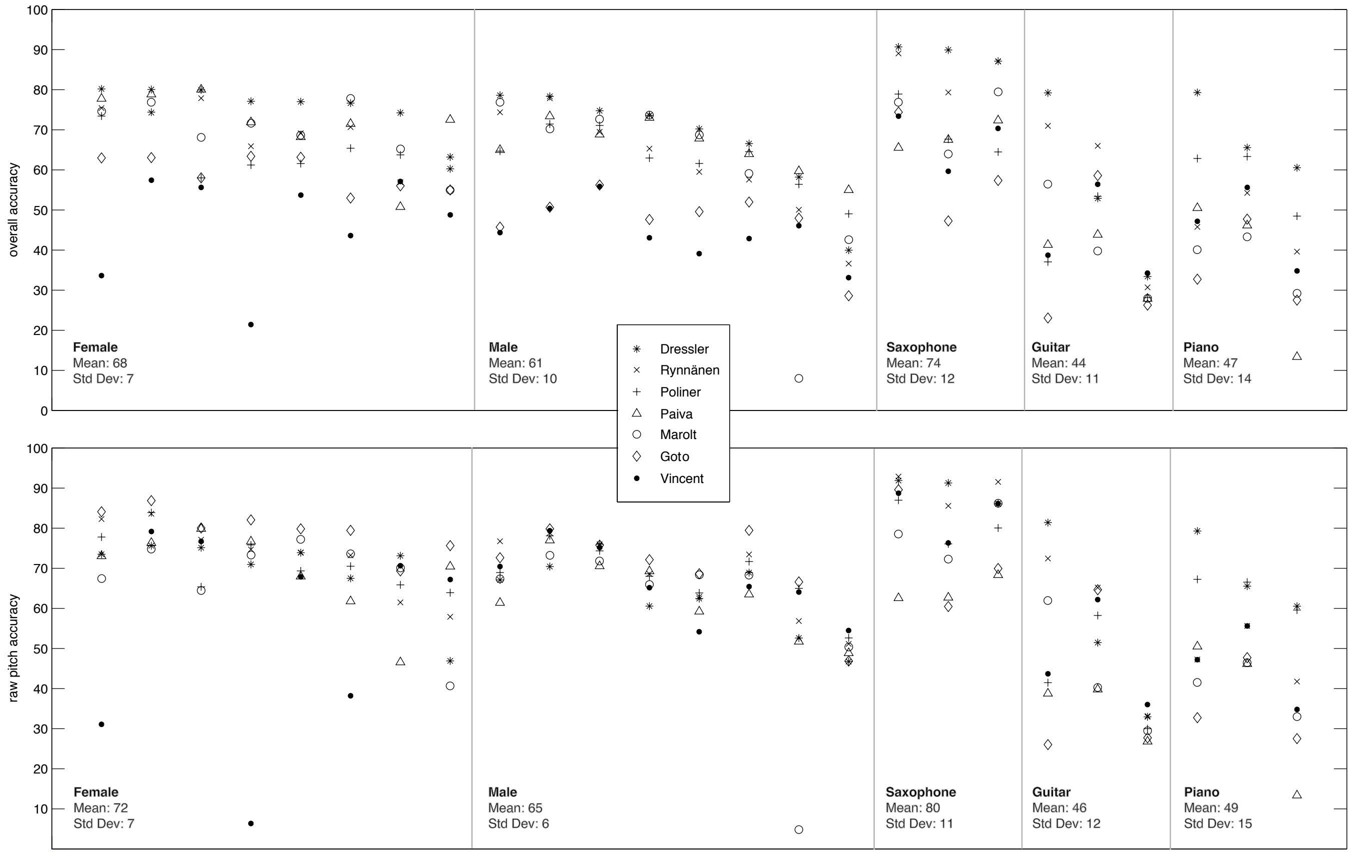

Fig. 7. Song-level overall accuracy and raw pitch accuracy for each algorithm across each of the individual pieces in the 2005 test set, ordered by lead instrument type and by relative difficulty.

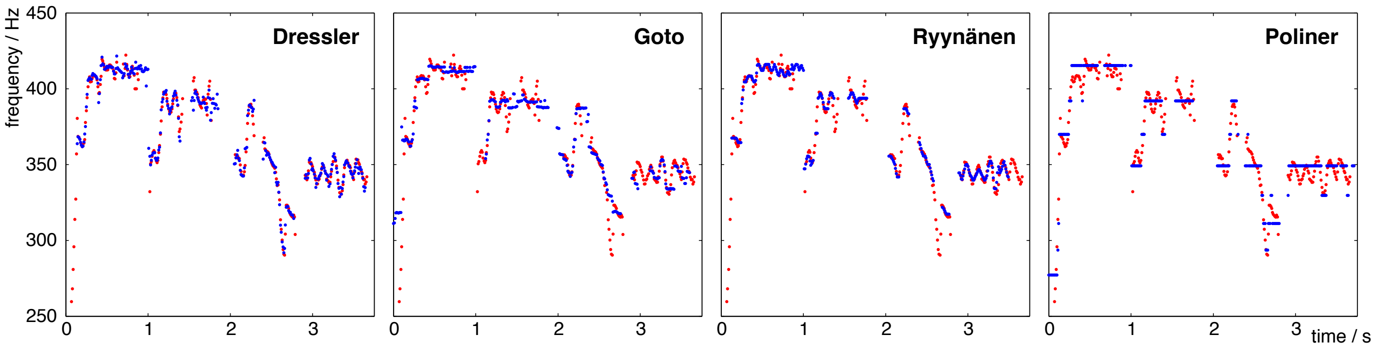

Fig. 8. Examples of actual melodies extracted by several of the submitted systems, compared to the ground truth (light dots) for a 3.7 s excerpt from the Madonna track.

Acknowledgments

This work was supported by the National Science Foundation (IIS-0238301, IIS-0340597 and IIS-0327371), by the Columbia AQF, by the European Commission through the SIMAC project (IST-FP6-507142), and by the Andrew W. Mellon Foundation.

References

[1] P. Cano. Fundamental frequency estimation in the SMS analysis. In *Proc. COST G6 Conference on Digital Audio Effects DAFx-98*, Barcelona, 1998.

[2] R. B. Dannenberg, W. P. Birmingham, G. Tzanetakis, C. Meek, N. Hu, and B. Pardo. The MUSART testbed for Query-by-Humming evaluation. In *Proc. Int. Conf. on Music Info. Retrieval ISMIR-03*, pages 40–46, 2003.

[3] A. de Cheveigné. Pitch perception models. In *Pitch: Neural Coding and Perception*. Springer Verlag, New York, 2005.

[4] A. de Cheveigne and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal Acoustic Society of America*, 111(4):1917–1930, 2002.

[5] J. Downie, K. West, A. Ehmann, and E. Vincent. The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview. In *Proc. 6th International Symposium on Music Information Retrieval ISMIR*, pages 320–323, London, 2005.

[6] K. Dressler. Extraction of the melody pitch contour from polyphonic audio. In *MIREX Audio Melody Extraction Contest Abstracts*, London,

2005.

[7] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2001.

[8] H. Duifhuis, L. F. Willems, and R. J. Sluyter. Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *J. Acoust. Soc. Am.*, 71(6):1568–1580, 1982.

[9] J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell System Tech. J.*, 45:1493–1509, 1966.

[10] A. Ghias, J. Logan, D. Chamberlin, and B. Smith. Query by humming – Music information retrieval in multimedia databases. In *Proc. ACM Multimedia*, San Francisco, CA, 1995.

[11] J. L. Goldstein. An optimum processor for the central formation of pitch of complex tones. *J. Acoust. Soc. Am.*, 54:1496–1516, 1973.

[12] E. Gomez, S. Streich, B. Ong, R. Paiva, S. Tappert, J. Batke, G. Poliner, D. Ellis, and J. Bello. A quantitative comparison of different approaches for melody extraction from polyphonic audio recordings. Technical Report MTG-TR-2006-01, Univ. Pompeu Fabra Music Tech. Group, 2006.

[13] M. Goto. A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, 2004.

[14] M. Goto. A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, 2004.

[15] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. Rwc music database: Popular, classical, and jazz music databases. In M. Fingerhut, editor, *Proc. Third International Conference on Music Information Retrieval ISMIR-02*, Paris, 2002. IRCAM.

[16] M. Goto and S. Hayamizu. A real-time music scene description system: Detecting melody and bass lines in audio signals. In *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pages 31–40, Stockholm, August 1999.

[17] D. J. Hermes. Pitch analysis. In *Visual Representations of Speech Signals*, page 325. Wiley, New York, 1993.

[18] W. Hess. *Pitch Determination of Speech Signals*. Springer-Verlag, Berlin, 1983.

[19] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Application of bayesian probability network to music scene analysis. In *Proceedings of the 1995 Int. Joint Conf. on AI Workshop on Computational Auditory Scene Analysis*, pages 52–59, Montreal, 1995.

[20] A. Klapuri. Multiple fundamental frequency estimation by harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Processing*, 11(6):804–816, 2003.

[21] R. C. Maher. Evaluation of a method for separating digitized duet signals. *Journal of the Audio Engineering Society*, 38(12):956–979, 1990.

[22] M. Marolt. A connectionist approach to transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6(3):439–449, 2004.

[23] M. Marolt. On finding melodic lines in audio recordings. In *Proc. DAFX*, Naples, 2004.

[24] J. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, 1(4):32–38, 1977.

[25] G. Ozcan, C. Isikhan, and A. Alpkocak. Melody extraction on midi music files. In *Proc. Seventh IEEE International Symposium on Multimedia (ISM'05)*, pages 414–422, 2005.

[26] R. P. Paiva, T. Mendes, and A. Cardoso. A methodology for detection of melody in polyphonic music signals. In *116th AES Convention*, 2004.

[27] G. Poliner and D. Ellis. A classification approach to melody transcription. In *Proc. Int. Conf. on Music Info. Retrieval ISMIR-05*, London, 2005.

[28] C. Raphael. Automatic transcription of piano music. In *Proc. Int. Conf. on Music Info. Retrieval ISMIR-02*, 2002.

[29] M. P. Ryynanen and A. Klapuri. Polyphonic music transcription using note event modeling. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY, 2005.

[30] H.-H. Shih, S. Narayanan, and C.-C. Kuo. Automatic main melody extraction from midi files with a modified lempel-ziv algorithm. In *Proc. ISIMVSP*, 2001.

[31] K. Sjölander and J. Beskow. WaveSurfer - an open source speech tool. In *Proc. Int. Conf. on Spoken Language Processing*, 2000.

[32] M. Slaney and R. F. Lyon. On the importance of time – A temporal representation of sound. In M. Cooke, S. Beet, and M. Crawford, editors, *Visual Representations of Speech Signals*. J. Wiley, 1993.

[33] D. Talkin. A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, chapter 14, pages 495–518. Elsevier, Amsterdam, 1995.

[34] T. Viitaniemi, A. Klapuri, and A. Eronen. A probabilistic model for the transcription of single-voice melodies. In *Proc. Finnish Signal Processing Symposium*, pages 5963–5957, 2003.

[35] E. Vincent and M. Plumbley. Predominant-f0 estimation using Bayesian harmonic waveform models. In *MIREX Audio Melody Extraction Contest Abstracts*, London, 2005.