

Christopher Sutton

Transcription of Vocal Melodies in Popular Music

MSc in Digital Music Processing
Queen Mary, University of London
2006

In undertaking this project I have been particularly fortunate to have not one but three supervisors, whose advice and encouragement have been a great help. My sincere thanks go to Emmanuel Vincent, Juan Bello-Correa and Mark Plumbley.

This report is submitted as part requirement for the degree of MSc in Digital Music Processing at the University of London. It is the product of my own labour except where indicated in the text. The report may be freely copied and distributed provided the source is acknowledged.

ABSTRACT

This report details the design, implementation and evaluation of a system for the transcription of vocal melodies in polyphonic recordings of popular music. The system operates on monaural sampled audio to produce a transcription indicating which time frames contain the vocal melody and for those that do, an estimation of its pitch.

A novel framework for transcription systems is proposed, in which multiple pitch estimation techniques are used in parallel and a modified Hidden Markov Model (HMM) is used to combine the estimates and produce a single transcription. This is shown to produce more accurate transcriptions than the individual pitch estimation techniques.

Two pitch estimation systems are presented, each tailored to the task of vocal transcription in a polyphonic context.

The first builds on a technique previously used for vocal detection, modifying it to produce a “semitone-cancellation” procedure which exploits the pitch instability of the human voice to attenuate non-vocal notes. A standard monophonic transcription method is then applied to produce pitch estimates.

The second is based on an investigation of the “high-frequency dominance” of the human singing voice which indicated that pitch estimates from the high frequency channels of a correlogram should favour vocal pitches. A correlogram pitch estimation system with high frequency bias is therefore used to provide a second set of pitch estimates.

The HMM used to combine estimates differs from standard post-processing HMMs in three main ways. Firstly, it takes estimates from multiple pitch estimation systems as its input, along with reliability measures which are used to choose between pitch candidates. Secondly, it makes use of a novel approach to voiced/unvoiced segmentation, in which the model infers silence based on the scattering of estimates from the two vocal-specific pitch estimation systems during vocal silence. Finally, unlike previous HMMs for post-processing of pitch estimates (which are restricted to semitone frequencies), the proposed model is designed to permit a continuous range of pitches which can better represent the pitch contours of the human voice.

The system is tested on a wide variety of popular music styles and instrumentations, and is found to exhibit similar pitch estimation accuracy to state of the art melody transcription systems. The novel approach to voiced/unvoiced segmentation also shows promising accuracy.

Table Of Contents

1	Introduction	1
1.1	Project aims	1
1.2	Automatic music transcription	1
1.3	Melody transcription	2
1.4	A system for vocal melody transcription	2
2	Background	4
2.1	Automatic transcription of sampled audio	4
2.2	Local pitch estimation	5
2.2.1	Two-way mismatch algorithm	5
2.2.2	Autocorrelation-based approaches	6
2.3	Hidden Markov Models for note modelling	8
2.4	Melody transcription	11
2.4.1	MIREX Melody Extraction competition	12
2.4.2	Pitch estimation techniques in MIREX 2005	12
2.4.3	Voiced/Unvoiced segmentation in MIREX 2005	14
2.4.4	Results of MIREX Melody Extraction competition	14
2.5	Characteristics of the human singing voice	15
2.5.1	Inverse comb filtering for vocal detection	18
2.5.2	Correlogram with high-frequency bias for vocal transcription	18
3	Design and Implementation	20
3.1	Overall system design	20
3.2	Semitone Cancellation and pitch estimation by Two-Way Mismatch	21
3.2.1	Semitone cancellation algorithm	22
3.2.2	Pitch estimation by TWM	23

3.2.3	Reliability measure	25
3.3	High-Frequency Correlogram	26
3.3.1	High frequency dominance experiments	26
3.3.2	Design of high-frequency correlogram	28
3.3.3	Combination of channel estimates	30
3.3.4	Reliability measure	31
3.4	Modified Hidden Markov Model	31
3.4.1	Dynamic state generation for continuous pitch	32
3.4.2	Observation probabilities based on all observed estimates	33
3.4.3	Transition probabilities and V/UV segmentation	36
3.4.4	Viterbi algorithm for state sequence inference	37
4	Evaluation	39
4.1	Test Set	39
4.1.1	Wiener filtering on Karaoke CDs	40
4.1.2	Multi-track data	41
4.1.3	Transcription of ground-truth melody	41
4.1.4	Combination of vocal and accompaniment recordings	41
4.2	Results	42
4.2.1	Informal listening tests for semitone-cancellation	42
4.2.2	Accuracy of pitch estimates	44
4.2.3	Accuracy of pitch estimates with HMM post-processing	44
4.2.4	Accuracy of proposed system	45
4.2.5	Comparison with existing systems	48
4.3	Future Work	49
5	Conclusions	51
	References	53

1. INTRODUCTION

1.1 Project aims

The aim of this project was to design and build a system capable of producing a clear representation of the melody of the main vocal part in a piece of popular music. The system should function accurately across a wide range of popular music genres and instrumentations. The task is born out of the more general melody transcription task, and if narrowing the aim has the desired effect, the resulting system should outperform general melody transcription systems when applied to music in which the melody is sung.

The motivation for designing such a system is explained below and a brief summary of the proposed system is given.

1.2 Automatic music transcription

The huge growth of digital music in recent years has led to a large number of musical recordings becoming available in digital form as sampled audio. Additionally, progress in electronic music production has resulted in a lot of symbolic music data being created (for example, MIDI[1] scores). However, for the most part these two exist in isolation. Sampled audio cannot be manipulated as easily as symbolic music formats, and symbolic formats lack the authenticity of real recordings.

A key step towards combining the benefits of these two realms is the ability to automatically produce a symbolic representation of a sampled music recording. This process is referred to as *musical audio transcription* and in its most general form is an unsolved problem in the research community[2][3][4]. This is primarily due to the complexity of an average music recording and the number of sound sources which interact and combine to produce only one or two channels of audio.

One way to make the problem more tractable is to restrict the task to audio recordings of a single instrument playing only one note at once — a *monophonic* recording (by contrast with *polyphonic* recordings which cover multiple simultaneous notes and/or multiple instruments). Such *monophonic transcription* is a much more manageable task and there exist various techniques which can accurately transcribe monophonic recordings (eg. YIN[5], TWM[6], and the correlogram[7]).

As a result of this success in monophonic transcription, researchers have tried to apply similar techniques to polyphonic recordings. One approach is to adapt the techniques to handle multiple simultaneous notes, and then to group the resulting note objects by instrument (“*streaming*”). Another method is to perform *source separation* to produce a monophonic recording for each instrument in the polyphonic mix, each of which can then each be handled by a monophonic transcription system.

A third possibility is to limit the extent of the transcription required. Rather than requiring the system to produce a symbolic representation of what each instrument plays, one can design a system to transcribe only a single instrument in the mix. This approach has been successful for the task of drum transcription for example[8]. It has also led to research into the task of *melody transcription*.

1.3 Melody transcription

Almost every piece of Western music has a prominent *melody* which is a key characteristic of the piece, and a natural focus for the human listener. It can be defined as the notes played by the “lead instrument”[9]. What determines which instrument is the “lead” at a given time in the music is not well defined and can be hard to determine in an automatic system. Nevertheless, the task of melody transcription is an attractive goal, as some of the tasks associated with music transcription could likely be tackled well using melody transcription alone.

For example, the classic “query by humming” task, in which a music recording is identified based on a short part of it being sung by a human. In most cases, this task would only require that the *melody* of the piece is held in a symbolic form, since that is the part of the piece most likely to be chosen by the human searcher. A full transcription of the music is probably not necessary.

Melody transcription is a relatively recent topic of research but there has been some success. Most notably, the Melody Extraction task in the Music Information Retrieval Evaluation eXchange (MIREX)[10] competition in 2004 and 2005 invited researchers to submit their melody transcription systems to be tested on a common set of data. The results were promising, but the transcription accuracy was considerably lower than is possible with monophonic transcription, and higher accuracy would be required for tasks such as that mentioned above.

1.4 A system for vocal melody transcription

The systems entered for the Melody Extraction competition displayed a variety of approaches to identifying the lead instrument. Some performed *predominant fundamental frequency (“F0”)* identification, transcribing the most salient pitch in each time frame. Others essentially performed polyphonic transcription and then used perceptual cues or grouping rules to identify which parts of the full transcription belonged to the lead instrument. Although detailed results are not yet available for the competition, the low scores for raw pitch accuracy suggest that failure to reliably identify the lead instrument is a large source of errors among systems.

The system proposed in this report was designed with a clear aim: to avoid such ambiguities in melody transcription by specifying that the vocal line carries the melody. The lead vocal line is a more tangible and precise target for transcription than the ill-defined and subjective concept of a “lead” instrument, and so by making this restriction in the problem definition, it is hoped that the task will be better defined and hence more tractable. This idea is supported by a paper by Li and

Wang[11] which shows that a system designed specifically for vocal melodies can more accurately transcribe melodies which are sung than a generic predominant F0 estimator.

This narrowed focus shouldn't incur a cost of generality in the applicability of the resulting system, since in the vast majority of popular music it will indeed be the vocal line which carries the melody.

The proposed system consists of three main subsystems. Unlike standard transcription approaches which use a single pitch estimation method, the proposed system uses two distinct pitch estimation systems. A modified Hidden Markov Model (HMM)[12] is used to combine the pitch estimates and make voicing decisions, producing as its output the final transcription.

The first pitch estimation system applies a Semitone-Cancellation procedure to attenuate accompaniment and leave vocal notes intact, and then uses a standard Two-Way Mismatch[6] pitch estimation method to produce pitch estimates and associated reliability measures. The Semitone-Cancellation procedure is based on the "Inverse Comb Filtering" technique used by Shenoy, Wu and Wang[13] to perform singing voice detection, but has been adapted to leave the vocal pitch intact and allow subsequent pitch estimation.

The second pitch estimation system uses a correlogram method based on a gammatone filterbank[7], modified for high-frequency bias by using only channels in the range 3-15kHz. This produces 19 pitch estimates each frame and a simple clustering procedure is used to choose a single pitch estimate and calculate an associated reliability measure. This approach is inspired by a system by Li and Wang[11] and supported by experiments carried out to verify the high-frequency dominance of the human singing voice.

The modified HMM system takes as input the two sets of pitch estimates and reliability measures. The reliability measures are used to select between pitch estimates and make voicing decisions. The system is designed to allow a continuous range of pitches, rather than quantising to semitone frequencies as is common among HMM-based transcription systems. The output from the HMM system is a sequence of pitch estimates at 10ms intervals, with estimates of 0 Hertz indicating that no vocal melody is present. This is the final output of the overall system for vocal melody transcription.

Background on the techniques used in the proposed system and the current state of the art in melody transcription is provided in Chapter 2. The design and implementation of the proposed system are covered by Chapter 3, detailing the two pitch estimation systems, and the modified HMM system in turn. The performance of the proposed system was evaluated on unseen data, and Chapter 4 describes the data set assembled for testing and the tests carried out. The results are also presented, along with some discussion and ideas for future work. Finally, Chapter 5 provides the conclusions of the report.

2. BACKGROUND

This chapter provides information on a number of areas related to the system proposed in this report. First the general task of automatic transcription of music is introduced (Section 2.1) and two approaches to local pitch estimation are described (Section 2.2). Next the use of Hidden Markov Models[12] for post-processing of pitch estimates is discussed in Section 2.3. The motivation for the task of melody transcription is introduced in Section 2.4, and the state of the art in melody transcription is discussed. Since the proposed system aims to transcribe vocal melodies only, the characteristics of the human voice which the system exploits are presented in Section 2.5.

2.1 Automatic transcription of sampled audio

One of the key goals in digital music research is to automatically produce high-level descriptions of the content of sampled musical audio. Such descriptions might include such information as the instruments playing, the artists performing, the musical style, the tempo and the actual score of the music played. This last possibility leads to the desire to perform automatic *transcription* of musical audio and produce a symbolic score of the music played. This could then be used for a wide variety of applications, such as resynthesis (possibly after modification), grouping of similar recordings, or intuitive searching of large music databases.

A fundamental distinction generally made when discussing transcription is between recordings and transcription systems which are *monophonic* and those which are *polyphonic*. The term *monophonic* implies that only a single note is played at once, and it is typically assumed that a single instrument is involved. The term *polyphonic* by contrast refers to situations where multiple notes may sound simultaneously, possibly played by multiple instruments. For example, a solo flute piece would be monophonic, while a solo piano piece would typically be polyphonic, as would a performance by a string quartet.

Although the polyphonic transcription task is generally much more challenging than the monophonic case, a large number of techniques first designed for the monophonic case have been adapted to handle polyphony and both styles of transcription have since been applied to the task of melody transcription.

The field of automatic transcription research is large and varied, and regrettably it is not possible here to give even a brief overview of all aspects of the topic. Whole areas such as streaming, onset detection, object coding, and auditory scene analysis are not even touched on, and those areas which are discussed (namely pitch detection and note modelling) are restricted to those techniques which are directly relevant to this project.

2.2 Local pitch estimation

In its simplest form, a transcription system consists of a method for *pitch estimation*. That is, a technique for determining the pitch or pitches present in each time frame. The *pitch* of a sound is intrinsically a perceptual phenomenon which is not necessarily easy to relate to the spectrum of the sound. In the simple case of a sinusoidal tone, the perceived pitch will be the frequency of the tone. Sounds made up of multiple sinusoidal components (or *partials*) will be perceived as having a strong pitch if they are *harmonic* (all partials are multiples of a single *fundamental frequency*), or slightly *inharmonic* (the partials are not exact multiples of a single frequency, but do follow a relationship based on a particular fundamental frequency).

In general the perceived pitch will be the fundamental frequency of the partials, but the relative strengths of the partials can influence the perceived pitch. For example, if the even partials of a sound (ie. those at even multiples of the fundamental frequency) are much stronger than the odd partials, the sound may be perceived as being an octave lower, as its fundamental frequency seems to be half the true value.

The notes produced by pitched musical instruments consist of a sum of sinusoidal components and will generally be harmonic or slightly inharmonic. Pitch estimation systems can therefore exploit assumptions about the spectral nature of musical notes to infer pitch from the frequency components present. It is common among pitch estimation methods to simply assume that note spectra are approximately harmonic.

The two pitch estimation methods discussed below are designed for monophonic contexts but may also be useful for predominant F0 estimation in a polyphonic context.

2.2.1 Two-way mismatch algorithm

The view of pitch presented above (as resulting from a harmonic relationship between sinusoidal partials) is equivalent to saying that a pitched sound has equally-spaced peaks in its power spectrum. This leads to techniques such as the “Two-Way Mismatch” algorithm proposed by Maher and Beauchamp[6] which compares the set of observed spectral peaks with the set of peaks hypothesised for a perfectly harmonic sound of a given fundamental frequency.

For each fundamental frequency being considered, a two-way mismatch (TWM) error is computed, by first comparing each observed peak with its closest hypothesis peak, and then comparing each hypothesis peak with its closest observed peak (see Figure 2.1). This accounts for the fact that there will often be predicted peaks which are not observed in the sound spectrum (due to the peculiarities of the instrument, or interference from other sounds), and there will often be observed peaks which a harmonic sound alone would not produce (primarily due to other sounds or background noise being present in the recording). The error calculation for each predicted–observed pair of closest peaks takes into account the frequency difference between the two, the amplitude of the observed peak, and the maximum amplitude of all observed peaks.

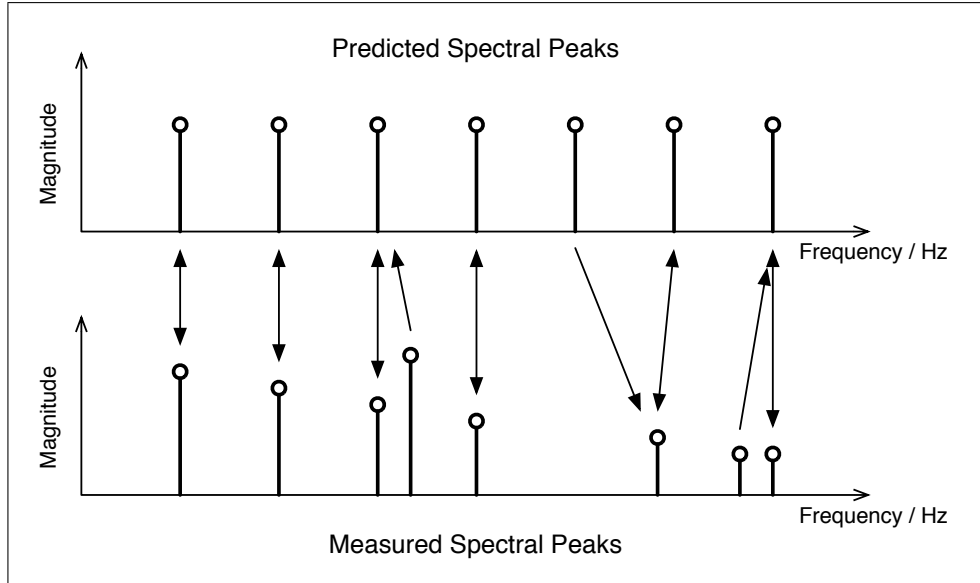


Figure 2.1: Illustration of the comparisons made during the two-way mismatch error calculation. Arrows indicate selection of the closest peak.

Maher and Beauchamp propose testing regularly spaced fundamental frequencies in a given range to find the frequency which produces the minimum TWM error. An implementation in [14] which was used in the system proposed in this report reduces the number of candidate pitches to be tested by considering only strong observed peaks in the spectrum and their octave-related counterparts.

2.2.2 Autocorrelation-based approaches

An alternative view of harmonic sounds is to consider their time domain behaviour. A sound composed of a number of sinusoidal components will exhibit strong periodicities in the time domain. One way to analyse such periodicities is to compute the *auto-correlation function* (or *ACF*) of the audio signal[5]. The ACF is defined by :

$$r(\tau) = \frac{1}{T} \sum_{t=0}^{T-\tau-1} x(t)x(t+\tau) \quad (2.1)$$

where x is the signal to be analysed between $t = 0$ and $t = T - 1$, and τ is the *lag* or period of repetition to be analysed.

When the ACF is computed in this way for finite windows of data, it can be seen to steadily decrease with lag due to the decreasing number of available data samples which are spaced by higher lags. It is useful to define an “unbiased” ACF as :

$$r(\tau) = \frac{1}{T - |\tau|} \sum_{t=0}^{T-\tau-1} x(t)x(t+\tau) \quad (2.2)$$

which compensates for this systematic bias. It is worth noting that there is still a problem at high lags, as the ACF values are weakly supported by the data and will

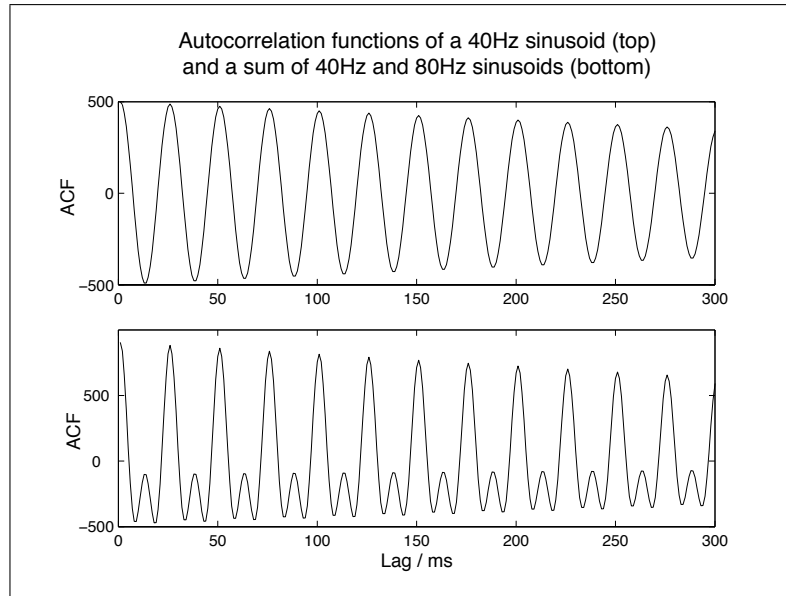


Figure 2.2: Illustration of the peaks of an autocorrelation function of a single sinusoid (top), and a pair of harmonically-related sinusoids (bottom)

be less reliable estimates of the true autocorrelation than those for small lags. To obtain reliable estimates, the autocorrelation function should be computed only for lags which are safely below the length of data window used.

A peak in the ACF at a particular lag indicates a strong periodicity at the frequency which is the reciprocal of the lag. So for example, if the ACF of an audio signal has a high peak at a lag of 25 milliseconds, the audio signal has a strong frequency component at 40 Hertz (Hz). Naturally a sound with a strong periodicity at 40 Hz will also exhibit strong periodicity at 20 Hz, 10 Hz, 5 Hz, and so on, resulting in a set of peaks in the ACF at lags which are multiples of that corresponding to the fundamental frequency. This is illustrated in the top plot of Figure 2.2. A harmonic sound will also exhibit peaks at lags which are fractions of the lag of the fundamental frequency, though typically these will be weaker peaks as the energy of partials tends to decrease as their frequency increases, and interference between harmonics will emphasise the fundamental frequency's peaks. This is illustrated in the bottom plot of Figure 2.2.

The fundamental frequency can therefore be estimated by consideration of the peaks of the ACF. A common approach is to assume the highest peak of the ACF corresponds to the fundamental frequency. Other approaches have also been successful, such as using multiple peaks to obtain a more accurate or reliable estimate of the fundamental frequency, for example by matching a harmonic comb template to the peaks[15].

Since most audio signals will have time-varying pitch content, it is standard to calculate the ACF for short frames of the signal, which may overlap. In this case the same equation is used with data from that frame only, to calculate the short-term ACF, and hence derive information about the local pitch of the signal.

With this basic ACF pitch estimator, it is difficult to exploit the information provided by higher partials than the fundamental, and no consideration is made of the human hearing system. Since pitch is a perceptual phenomenon, some consideration of human perception in pitch estimation systems can be helpful. A popular variant, the *correlogram*[16], therefore uses a perceptual model of the human hearing system which divides the signal into suitably spaced frequency bands and computes the ACF of each, allowing information from all bands to be considered when estimating the pitch.

The perceptual model used will vary between correlogram designs, but might include filtering to model the outer and middle ear, followed by a filterbank whose channels model the bandwidths of the critical bands of the human ear, and then modelling of the inner hair cells by half-wave rectification and lowpass filtering[7].

Even though a channel’s frequency range may be far above the range of fundamental frequencies under consideration, the lowpassed signal derived in this way contains information about the low frequency content of interest. This phenomenon is referred to as the “beating effect”, and occurs because the partials of a harmonic sound combine and interfere in such a way that the overall envelope of the sound fluctuates at the rate of the fundamental frequency. This means that the envelope of high-frequency channels fluctuates at the fundamental frequency of the sound, and so the ACF of the envelope of high-frequency channels will reflect the fundamental frequency, even though it is outside the frequency range of the channel.

Each filterbank channel therefore contributes information about the fundamental frequency. The time-domain signal from each filterbank channel can be analysed to derive its ACF using equation 2.1 or 2.2. Often these per-channel ACFs are combined by normalising (based on the channel energy) and summing them to produce a summary autocorrelation function, which could then be analysed to produce pitch estimates as described above.

Other approaches may be employed to combine channel estimates, such as a statistical approach[11], discussed in Section 2.5.2, or the approach taken in the system proposed here, which is to perform clustering of the estimates from a subset of channels (see Section 3.3).

2.3 Hidden Markov Models for note modelling

Although the methods described above may produce reliable pitch estimates in a monophonic transcription task, there will still be occasional errors, and estimates will be given even when there is no note being played. There is therefore a need for a post-processing system to produce a more musically accurate representation of the notes being played. This may involve rejecting spurious estimates in the middle of stable pitch regions, performing segmentation between note regions and silence (“*Voiced/Unvoiced*”, or “*V/UV*” segmentation), and quantising the exact pitches observed to the most likely semitone note being played (the Western pitch system contains twelve *semitones* or *half steps* in each octave). This results in a very clean

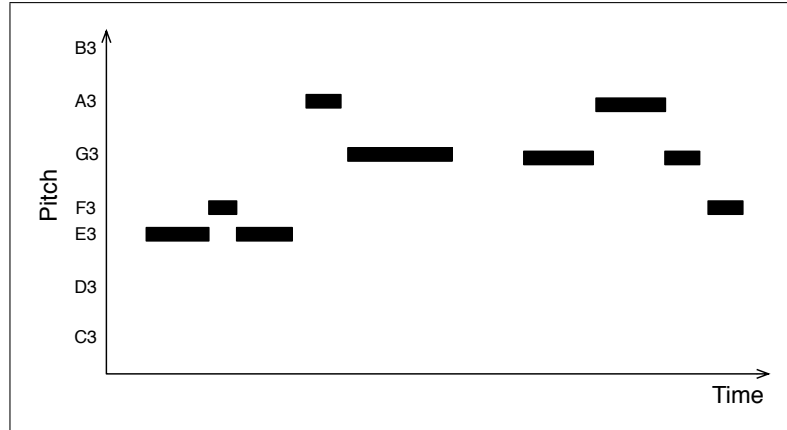


Figure 2.3: Illustration of a melody expressed in “piano roll” form.

“piano roll” type score, in which each note is represented by a period of constant pitch (see Figure 2.3 for an example).

A popular choice for this post-processing of pitch information is to use a Hidden Markov Model (HMM)[12], which is a probabilistic finite state machine in which the state occupations and transitions are not directly observed but may be inferred based on some observed data, and transition probabilities between states depend only on the current state. In the case of music transcription, the observed data is typically the output of the pitch estimation system, and states correspond to semitones of the Western chromatic scale. There exist efficient algorithms for training HMMs and for inferring the most likely state sequence given a set of observed data, most notably the Baum-Welch algorithm[12] for the first task and Viterbi[12] or Token-Passing[17] algorithms for the second task.

One example of a successful application of HMMs to transcription is the system introduced by Ryyänänen and Klapuri[18] which is designed to take observations of pitch, voicing, accent and meter and produce a piano-roll type score. The system was designed for transcription of singing voice, which (as is discussed further below in Section 2.5) exhibits strong fluctuations about the true note frequency. The HMM method they propose halved the number of transcription errors when compared with simply rounding each pitch estimate to the nearest semitone. A brief summary of the system is given here for the sake of later comparison with the HMM proposed by this report.

The system consists of two core models which are combined in a HMM framework. The first is a “Note Model”, itself a HMM, which has three states referred to in the paper as the “transient”, “sustain”, and “silence” stages of a note being played, and probability distributions for the four observed quantities (pitch, voicing, accent, meter) during each state. This captures behaviour such as that the difference between the nominal frequency of the note model and the observed frequency is likely to be small during the sustain stage, but large during transitions between notes. The voicing feature on the other hand, which is low for strongly-voiced estimates tends to be low during transient or sustain stages, but becomes bimodal with peaks at high and low values during the silence stage (as pitch estimates may or may not be reliable).

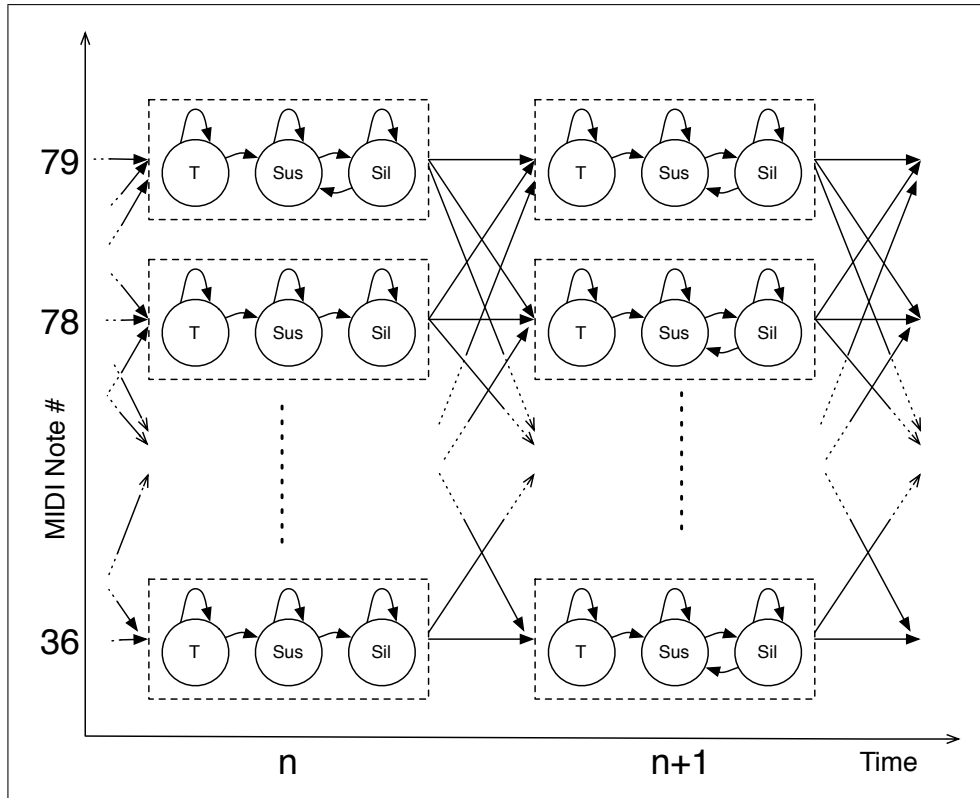


Figure 2.4: Illustration of states in Ryyänen and Klapuri's note modelling HMM.

These observation distributions not only influence the transitions between the three states of a given note, they indicate to the overall transcription HMM which notes are likely given the observed data. The overall transcription system uses 44 of these note models, one corresponding to each semitone in a likely range of sung notes. At each point in time therefore, one of three states of any one of the 44 notes may be occupied by the HMM (see Figure 2.4).

These note HMMs are linked over time by the second model in the system, the “Musicological Model”, which makes use of a key-estimation algorithm to determine the musical key of the music. In a given key, the occurrence of particular notes, and transitions between particular pairs of notes are much more likely than others. The transcription system exploits this fact to assign suitable transition probabilities between note HMMs based on probability distributions learned from a large melody database.

In a later system for polyphonic transcription[3] Ryyänen and Klapuri took multiple pitch estimates and computed the observation probability for each note model using the single estimate whose pitch was closest to the nominal frequency of that note model. The transcription was extracted by repeated application of the token-passing algorithm, with the chosen states from one pass being excluded for the next pass, to produce multiple disjoint note sequences. It also added a “silence model”, a single-state HMM whose observation probability was computed based on the largest observation probability of any state in that time frame. This allowed simple V/UV segmentation as part of the transcription, where silence will generally be chosen when no voiced state has strong observation probability.

For the sake of later comparison with the HMM proposed in this paper, it is important to note the following about the two systems[18][3] described above:

- A single, fixed set of states is defined and used for all time instants.
- The observation probability for each note state is computed based on a single observation from the pitch estimation system.
- The system outputs a piano-roll type score which represents the unstable pitch of the input sound as periods of constant pitch.
- Voiced/Unvoiced segmentation is performed by defining the probability of silence as $(1 - p)$, where p is the probability of the most likely voiced state.

2.4 Melody transcription

Research into polyphonic transcription has enjoyed some success in recent years, but the accuracy of transcription still does not rival that of the monophonic case (for example, a modern polyphonic transcription system[3] achieves around 40% recall and precision scores on notes, whereas a monophonic system by the same authors[18] transcribes over 90% of notes correctly). This has led some researchers to investigate producing a reduced transcription from a polyphonic recording, such as transcribing chords[19], the melody and bass lines[20], or indeed just the melody line[21][22][23].

The intention of narrowing the transcription aim in these ways is that the task may be made more manageable while still producing useful results. For example, focussing on a single melody and bass line means that monophonic transcription techniques may be applied after suitable pre-processing. At the same time, many transcription-dependent applications do not require a full transcription of every note played. For example, transcribing the melody could well suffice to implement novel and intuitive methods for searching large music databases.

Even with the definition of melody as “the notes played by the lead instrument” there is room for ambiguity. Often the term is used to refer to what is more accurately called “predominant fundamental frequency estimation”, in which the single most dominant pitch at every point in time is transcribed. Though this does not necessarily preclude V/UV segmentation and short periods of silence, such a system is likely to transcribe the notes of several different instruments over a short time interval, depending on which is deemed to have the predominant pitch in each time frame. This does not fit well with the intuitive notion of a melody.

Another way to interpret the term “melody” is to try to identify how humans decide which notes make up the melody in a polyphonic arrangement, and use cues based on these notions to identify parts of a polyphonic transcription which could be the melody. This might also involve instrument tracking once an instrument is identified as carrying the melody. Similar ideas can be found in systems which use a perceptual model as part of their pitch estimation stage - for example, the PreFest

system[20] bandpasses a particular frequency range for transcription of the melody, as that is the range humans prefer to hear and write melodies in.

The task of lead instrument identification and tracking to perform melody transcription is challenging. Another possibility therefore may be to introduce clear restrictions on what constitutes a melody, which, though they do not always meet with the intuitive notion of a melody, can allow the task to be tackled well in a subset of cases. One such restriction is to assume a particular instrument carries the melody — in the case of this report, the assumption is that all melodies are sung.

Clearly to define “melody” as being something which is sung is musically unsound, but since in the vast majority of popular music the main melodies *are* sung, it is not so broad an assumption as it first appears. This is a re-application of the narrowing of focus which led to the task of melody transcription, and in much the same way, it may be that a narrowed task is more easily tackled and can still permit many of the same applications as the wider task.

It is worth noting that in the melody extraction competition discussed below in Section 2.4.1, none of the systems entered restricted their definition of melody to sung melodies, but in 9 out of 13 of the provided training samples[24] the ground truth melody was defined as being the pitch contour of the lead vocals.

2.4.1 MIREX Melody Extraction competition

As part of the 2005 International Symposium on Music Information Retrieval (ISMIR) conference, the Music Information Retrieval Evaluation eXchange (MIREX)[10] ran a competition on Melody Extraction, in which eight systems were tested on a common data set with common evaluation criteria. This provides a good survey of current approaches to melody transcription, as well as providing information about the current standard of transcription accuracy. The discussion on methods for pitch estimation and V/UV segmentation in melody extraction therefore focuses on those eight systems.

The systems were tested on 25 excerpts from a range of musical genres, each 10–40 seconds long. Each excerpt featured a single lead instrument throughout, and the pitch of that instrument defined the ground truth melody. The systems were scored on their accuracy of V/UV segmentation and pitch estimation, and given an overall accuracy score incorporating both. Some summary results are given below, but the full evaluation results from the competition can be found on the competition webpage[25] and in some of the papers cited below describing the systems entered.

2.4.2 Pitch estimation techniques in MIREX 2005

The particular pitch estimation algorithms chosen for each system vary considerably and are probably not as important to the performance of the system as the decisions made about how to identify the melody. The approaches to identifying

System	Approach	V/UV segmentation method	Pitch Output
Brossier[26]	Predominant F0	-	Exact
Dressler[21]	Streaming	No suitable stream \Rightarrow silence	Exact
Goto[27]	Predominant F0	-	Exact
Marolt[28]	Streaming	No suitable stream \Rightarrow silence	Exact
Paiva[29]	Streaming	No suitable stream \Rightarrow silence	Quantised
Poliner & Ellis[23]	Classification	Energy threshold	Quantised
Ryynänen & Klapuri[22]	Polyphonic HMM	Silence model in HMM	Exact
Vincent & Plumbley[30]	Predominant F0	-	Exact

Table 2.1: Summary of approaches taken in the 2005 MIREX Melody Extraction competition.

melody fall into four groups: predominant F0, pure classification, streaming, and polyphonic HMM-based.

Three of the systems entered (Brossier[26], Goto[27], Vincent and Plumbley[30]) are predominant fundamental frequency estimation systems, defining the melody by the most salient note in each time frame.

Brossier’s system[26] filtered the input to emphasise mid-range frequencies, and then used a phase vocoder and harmonic comb matching system to identify the most likely fundamental frequency in each frame. Goto’s system[27] also filters to select a suitable frequency range for the melody, and then performs a probabilistic Maximum *A Posteriori* Expectation Maximisation using a harmonic model of note spectra to infer the predominant fundamental frequency for each frame. These can both be seen as predominant fundamental frequency estimation methods within a particular frequency range. The system by Vincent and Plumbley[30] used a modified YIN[5] approach to select some pitch candidates and a Bayesian waveform modelling algorithm to select the predominant F0 from among them.

The system built by Poliner and Ellis[23] stands apart from the rest of the systems entered in that it is a pure classification approach in which spectrum data is input to a trained classifier which then outputs a single pitch per time frame. It is therefore difficult to say what exactly defines the melody in this system.

The other four systems take a more polyphonic approach, making use of multiple note candidates in each frame. Three of the systems, those of Dressler[21], Paiva[29] and Marolt[28] go so far as to group pitch candidates into “tracks” or “streams” which are fragments of pitch contours, and these fragments are then evaluated based on rules of human perception, clustering by similar features, or measures of salience in order to choose which will make up the final melody transcription.

The system entered by Ryynänen and Klapuri[22] is essentially the polyphonic transcription system described in Section 2.3 above, except that only a single pitch track is built by the note modelling HMM (in the full polyphonic transcription system other parts are transcribed by subsequent token-passing algorithm runs).

The melody is therefore defined as the most probable path through the note HMM based on pitch estimates for the whole polyphonic mix. A post-processing step of selecting the closest pitch candidate to each transcribed semitone frequency is used to produce an exact pitch transcription rather than the quantised output of the HMM.

Of the eight systems, six output exact frequency values each time frame, while two elected to quantise to nominal semitone frequencies (see Table 2.1). As will be seen below in Section 2.5, vibrato and portamento in vocal melodies can lead to frequency deviation by up to 200 cents from the nominal pitch, and the correctness criterion for pitching in the MIREX evaluation was that the estimate be within 50 cents of the ground truth (a *cent* is one percent of a semitone interval). As a result, systems which quantise their output may have suffered due to their choice of output format. However, since one of the three top performing systems in terms of pitch accuracy did output quantised pitches, it was probably not a large source of error.

2.4.3 Voiced/Unvoiced segmentation in MIREX 2005

Apart from pitch estimation, the key goal in melody transcription is to perform voiced/unvoiced segmentation, identifying which time frames contain a melody pitch and which should be marked “silent”. For systems which view melody as “predominant fundamental frequency”, all time frames can be considered voiced, and so the systems by Brossier[26], Goto[27] and Vincent and Plumbley[30] performed no V/UV segmentation.

The three systems (Dressler[21], Paiva[29] and Marolt[28]) which performed grouping of pitch candidates into streams or tracks all performed V/UV segmentation by defining silence to occur when no suitable note objects are present (with the definition of “suitable” depending on the particular features assigned to note tracks by each of the systems).

The classification system by Poliner and Ellis[23] used a simple energy threshold in the frequency range 200–1800 Hz, where silence is transcribed when the energy in a time frame falls below a certain threshold.

The note HMM system by Ryyänen and Klapuri[22] uses the previously mentioned silence model to transcribe silence based on the maximum observation probability across all possible states. Roughly speaking, if no note model is probable over a period, silence will be chosen instead.

2.4.4 Results of MIREX Melody Extraction competition

Evaluation of pitch accuracy was performed by calculating the percentage of voiced frames which were estimated within 50 cents of the ground truth pitch. It was possible to provide a pitch estimate even during frames transcribed as “silent”, so that pitch accuracy scores do not suffer as a result of poor V/UV segmentation. Systems scored in the range 60%–70%, with the top scoring system being the note modelling HMM by Ryyänen and Klapuri[22], closely followed by the systems of

Dressler[21], Poliner and Ellis[23], and Goto[27]. It is interesting to note that Goto’s PreFEst system which was designed between 1998 and 2000 is still one of the top scoring systems in this regard.

A common source of errors in transcription systems is to transcribe the note an octave above or below the true note — This happens because notes separated by an octave offset have fundamental frequencies which differ by factors of two, such that half the harmonics of a given note will overlap with those of the note an octave below. The MIREX evaluators therefore calculated an additional score for “chroma accuracy”, meaning the pitch estimation accuracy when octave errors are permitted. These scores were 3%–12% higher than the raw pitch accuracy scores, suggesting that octave errors were a significant source of transcription error in these systems. Even allowing octave errors, the scores for pitch accuracy are quite low. This suggests that systems were not able to reliably identify which notes belong to the lead instrument.

The voicing scores must be evaluated in a manner which is independent of the proportion of frames which are voiced or unvoiced, and so scores were given as the percentage of truly voiced frames labelled as voiced (the “hit rate”), the percentage of unvoiced frames labelled as voiced (the “false alarm rate”) and the d-prime measure[31] which gives an overall measure of the sensitivity of the detector. Values for the d-prime measure ranged from 0.83 to 1.85, where high values indicate clearer separation of voiced and unvoiced frames, and with the top scoring system ($d' = 1.85$) exhibiting a hit rate of 82% and a false alarm rate of 17%.

“Overall” scores were calculated as the percentage of frames correctly transcribed (that is, voiced frames are estimated within 50 cents of the ground truth pitch, unvoiced frames are labelled unvoiced), and scores ranged from 46%–71%. Notably the system with highest score is Dressler’s[21], which also exhibits the strongest d-prime measure for voicing. Ideally V/UV segmentation should be strong enough to improve upon the raw pitch accuracy scores, but only two of the systems showed higher overall scores than pitch scores. This suggests that improved voicing accuracy could go a long way towards improving overall transcription accuracy.

The competition showcased a number of promising approaches to melody transcription, but it is clear that further work is necessary to achieve the accuracies required by applications based on the technology. There is room for considerable improvement in both the pitch estimation and the voicing segmentation tasks. Research into melody transcription is ongoing, and the Melody Extraction task will be included in this year’s MIREX competition at the ISMIR 2006 conference.

2.5 Characteristics of the human singing voice

The work described in this report seeks to perform reliable melody transcription by exploiting characteristics of the human singing voice to identify vocal melodies. Some way of tailoring the melody transcription system to favour the human voice was required, and there are two vocal characteristics used by the proposed system : *pitch instability* and *high-frequency dominance*. These characteristics are discussed

here, and background on the methods chosen to exploit them is given in the next two sections.

The first characteristic, “pitch instability”, includes a number of factors which result in the pitch contour of the singing voice being strikingly different from most pitched instruments. An example can be seen in Figure 2.5, in which the pitch contour of melody excerpts for a voice, guitar, and piano are shown.

One factor in producing the constantly varying pitch of a sung note is vibrato — the roughly regular frequency modulation which occurs to some extent whether the singer is aware of it or not (eg. top plot of Figure 2.5, time frames 100–150). It is in the nature of the instrument that no pitch produced is perfectly stable, and though the production method of vibrato is not conclusively known[32], some studies have measured the typical rate (number of fluctuations per second) and extent (maximum displacement from the average pitch value) of vibrato. These numbers vary, mainly depending on the singer, but also across different notes sung by the same singer, or even during the course of a particular sung note. The most useful reported statistic as far as this report is concerned is from a study by Timmers and Desain[33], cited in [13], which reports that the human singing voice typically exhibits a vibrato extent of ± 60 –200 cents, while other instruments typically exhibit vibrato extents of ± 20 –35 cents.

There are several other factors involved in pitch instability. For example, vocalists almost always sing *legato*, changing pitch smoothly from one note to the next (eg. time frames 30–40 and 185–190 of the top plot of Figure 2.5), and even notes which are isolated or at the beginnings of phrases exhibit a smooth approach to the note from above or below the true pitch (eg. time frames 15–20 and 160–175 of the top plot of Figure 2.5).

The second characteristic, referred to here as “high-frequency dominance”, refers to the fact that compared with most pitched instruments, the voice has a lot of energy in its high frequency partials. This seems to have been exploited in the system described below in Section 2.5.2, and was investigated further as part of this project — the results may be found in Section 3.3.1. It was established that when a vocal and non-vocal sound with the same energy are compared, the vocal sound has more energy in high frequencies (specifically the 3kHz to 15kHz range) than the non-vocal sound. This can be exploited to distinguish between vocal notes and strong notes from other instruments.

Aside from these two characteristics, it was helpful in deciding suitable frequency ranges for parts of the proposed system to know that the frequency range of sung notes is roughly 80Hz–1000Hz[11].

There are two existing systems which seek to exploit the two characteristics described above. The first exploits pitch instability to perform vocal detection (Section 2.5.1), while the second seems to exploit high frequency dominance to perform transcription of vocal melodies (Section 2.5.2).

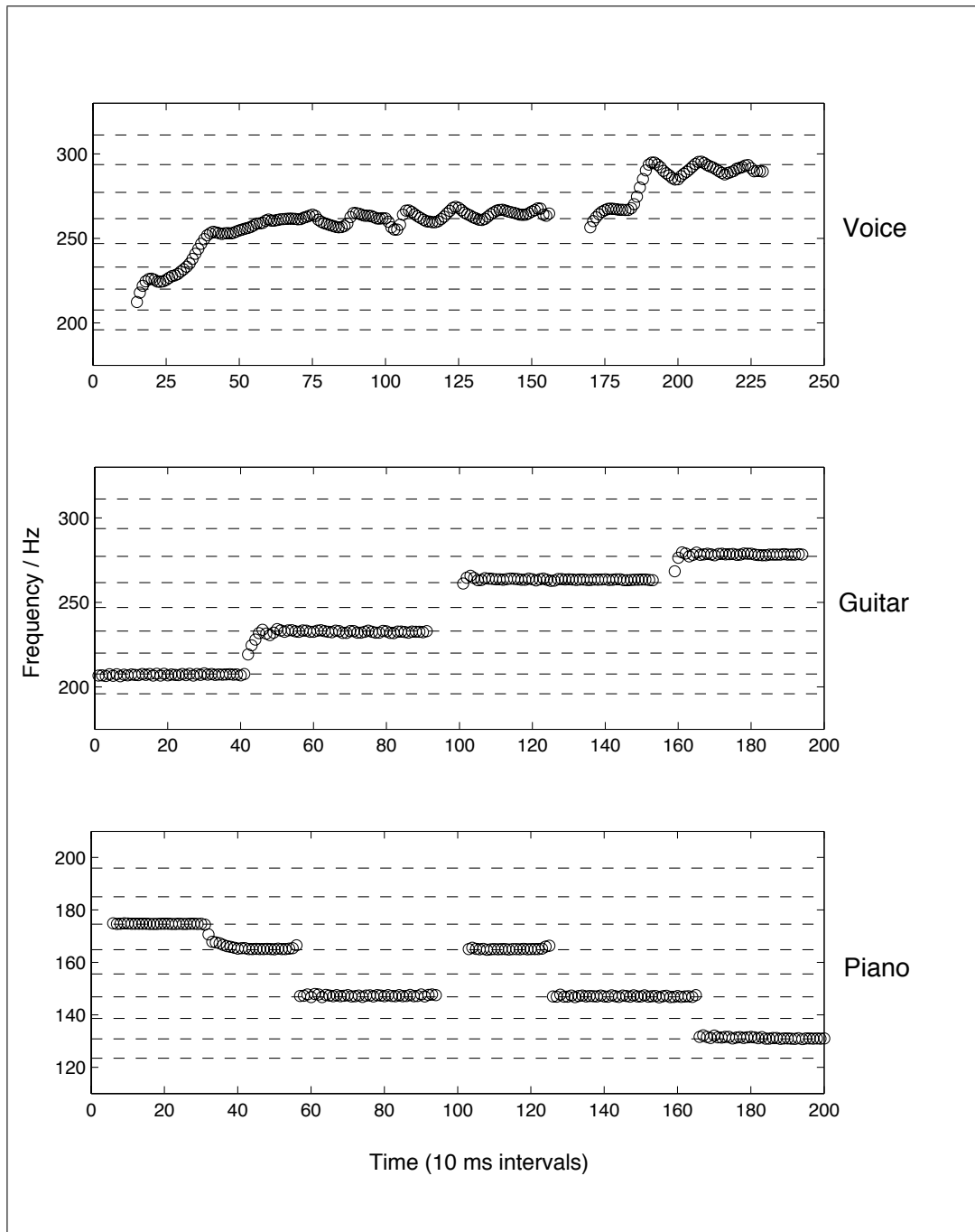


Figure 2.5: Examples of the pitch contour of the voice, guitar and piano (playing different melodies). Dashed lines show the fundamental frequencies of musical notes.

2.5.1 Inverse comb filtering for vocal detection

A paper by Shenoy, Wu and Wang[13] describes a system built to align lyrics with music for the purpose of generating karaoke displays automatically. One of the important aspects of this system is a V/UV segmentation algorithm for identifying time frames containing vocals.

The segmentation algorithm uses a technique called “Inverse Comb Filtering” to preserve the energy of sung notes while attenuating notes from other instruments. First, key detection is performed to identify which notes are likely. Then, for each of the seven notes in the key’s scale, the audio is filtered by an inverse comb filter which attenuates frequencies at integer multiples of the fundamental frequency of that note.

If all notes were pitched at the correct semitone frequencies and perfectly harmonic, and the inverse comb filters ideal, this process would remove all notes in the key of the song. In fact the filters will only attenuate the target frequencies to a certain degree, and most notes aren’t perfectly harmonic, and so won’t be fully attenuated. More importantly, not all notes are pitched at the ideal semitone frequency — in particular, while most instruments will be well-tuned and pitch notes accurately, exhibiting only a small degree of vibrato, sung notes (as was seen in section 2.5) have a pitch contour which is only occasionally at the exact target pitch.

The result is that while most instrumental notes are attenuated by the process, vocal notes survive relatively unharmed. Percussion (with its broadband spectrum) will also survive the process, but since percussive notes tend to be short-lived, their effect is localised in time. Most sustained energy therefore belongs to vocal notes, and the energy of the resulting signal can be suitably thresholded to provide vocal/non-vocal segmentation.

2.5.2 Correlogram with high-frequency bias for vocal transcription

The only existing system found which is specifically designed for transcribing the pitch of singing voice in polyphonic mixes is that of Li and Wang[11], which uses a correlogram approach. The system begins by downsampling the input audio to 16kHz and passing the signal through a 128-channel gammatone filterbank to model the critical bands of the human ear[34]. For each channel the ACF is calculated, either directly on the channel output (for “low” channels — below 800Hz) or on the envelope of the channel output (for “high” channels — above 800Hz).

The authors found experimentally that high frequency channels were reliable estimators of the singing voice pitch and could all be used, while only some low frequency channels were selected in each time frame, based on the strength of their first ACF peak. Pitch candidates were evaluated using a probabilistic model of reliable and unreliable channels’ ACFs, and a single pitch estimate produced per time frame.

The authors reference the “beating effect” (see Section 2.2.2) as being responsible for the high frequency channels providing reliable estimates of the fundamental frequency of the singing voice. Although the beating effect explains how high frequency channels may estimate a fundamental frequency lower than their own frequency range, it does not explain why the vocal pitch should be preferred — instrumental notes are often as loud or louder than the vocal melody, and it should therefore be those notes which are estimated via the beating effect. The stated reliability of high frequency channels as estimators of vocal pitch in this paper therefore led to the investigation of high-frequency dominance described in Section 3.3.1. This in turn led to the development of a modified correlogram approach to vocal pitch estimation, described in Section 3.3.

The evaluation provided by Li and Wang serves to demonstrate that the system identifies vocal pitch more reliably than predominant fundamental frequency systems, specifically the standard correlogram, Ryyänänen and Klapuri’s multi-pitch estimation method[3] and a system by Wu et al.[35]. The accuracy scores are calculated based on whether the estimated pitch is within 20% of the reference pitch — this is a very rough measure of transcription accuracy, but comparing such scores against the other systems demonstrates that this system does preferentially transcribe vocal pitches rather than just the predominant pitch.

3. DESIGN AND IMPLEMENTATION

This chapter details the design and implementation of the proposed vocal melody transcription system for popular music recordings. First, a brief overview of the system structure is given in Section 3.1. Since the design is quite different from previous approaches to melody transcription, some justification of the system structure is provided.

The system was designed to be modular and so the next sections each tackle a particular part of the system. Section 3.2 describes semitone cancellation pre-processing and the two-way mismatch pitch estimation system. Section 3.3 describes the high-frequency correlogram pitch estimation system. Section 3.4 describes the modified Hidden Markov Model system for combining pitch estimates and producing the final vocal melody transcription.

The system was implemented in Matlab for the sake of rapid prototyping and relatively efficient operation. During development a set of nine 15-second recordings and associated ground truth melody transcriptions provided for training in the MIREX 2005 competition[24] was used to test the pitch estimation systems, and later to train parameters for the modified HMM.

3.1 Overall system design

As was seen in Chapter 2, many transcription systems consist of a pitch estimation system, followed by some procedure for producing a more musical representation of the notes played (by removing erroneous pitch estimates, quantising pitch values to semitone frequencies, performing V/UV segmentation, etc.). Some systems use a pitch estimation system which produces multiple estimates for each time frame, all of which are input to the post-processing stage to produce either a single pitch track, or multiple tracks (in the case of polyphonic transcription).

This report proposes a novel approach in which multiple distinct pitch estimation techniques are used in parallel, and their estimates combined by a post-processing system to produce a single transcription, hopefully more accurate than using any one of the pitch estimation techniques. In this case, two pitch estimation systems were used, but the overall system design and the method for combining the estimates from each system generalise directly to a larger number of pitch estimation systems being used.

A high-level system design diagram can be found in Figure 3.1. The audio recording in the form of mono 16-bit wave data sampled at 44.1kHz (standard stereo CD format converted to mono) is input to the two pitch estimation systems. The two systems operate on this data (details in Sections 3.2 and 3.3 below) to produce a series of pitch estimates and associated *reliability measures* at 10 millisecond intervals. The reliability measures indicate the confidence of the pitch estimation systems

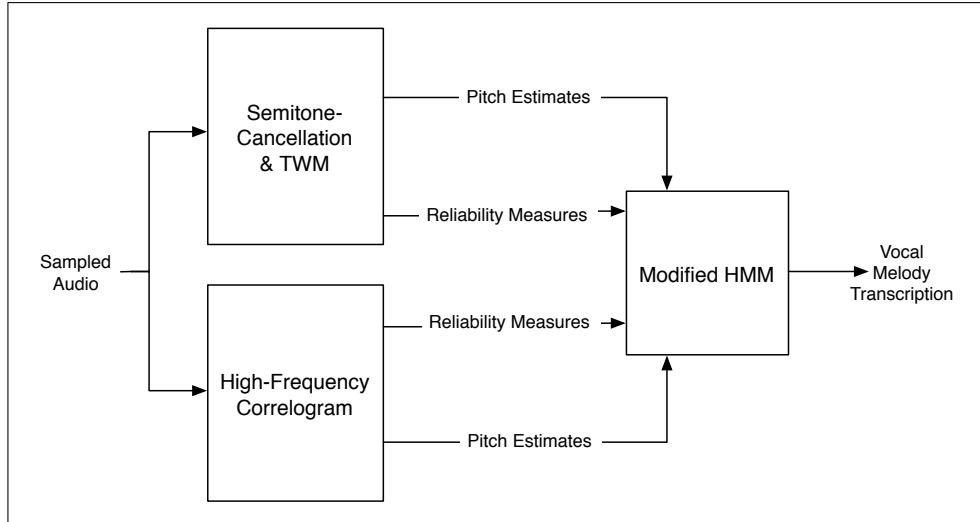


Figure 3.1: System Design Overview

in each of their estimates, and are an important part of combining the estimates from multiple systems. These pitch estimates and reliability measures are input to a modified HMM system (details in Section 3.4) which produces a single set of pitch estimates, with silence represented by estimates of zero Hertz. This output is the final transcription from the system.

3.2 Semitone Cancellation and pitch estimation by Two-Way Mismatch

The Semitone Cancellation-based pitch estimation system was designed in the vein of melody transcription systems which seek to isolate the melody and then perform monophonic transcription. It consists of a pre-processing stage in which a semitone cancellation algorithm based on the ideas of inverse comb filtering (introduced in Section 2.5.1) emphasises the vocals and attenuates other pitched instruments, followed by a standard frequency-domain monophonic transcription algorithm, the two-way mismatch algorithm (introduced in Section 2.2.1), the advantages of which are discussed below.

An overview of the design is provided in Figure 3.2. The TWM algorithm is not shown in detail since an existing implementation[14] was used without significant modification. The focus in this section is therefore on the pre-processing stage.

This pitch estimation system seeks to exploit the pitch instability property of the human singing voice to distinguish it from other pitched instruments. It was initially hoped that features of vocal pitch contours could be used to distinguish between partials belonging to the voice and those belonging to other instruments. For example, by applying a classifier algorithm to measured vibrato features of the pitch tracks found by SMS[36] analysis. Unfortunately, the difficulty of correctly linking spectral peaks into tracks corresponding to partials in a polyphonic context[37] makes it hard to exploit the features in this way.

An elegant solution to this problem is provided by Shenoy, Wu and Wang[13] and was introduced in Section 2.5.1. Rather than identify partials from all instru-

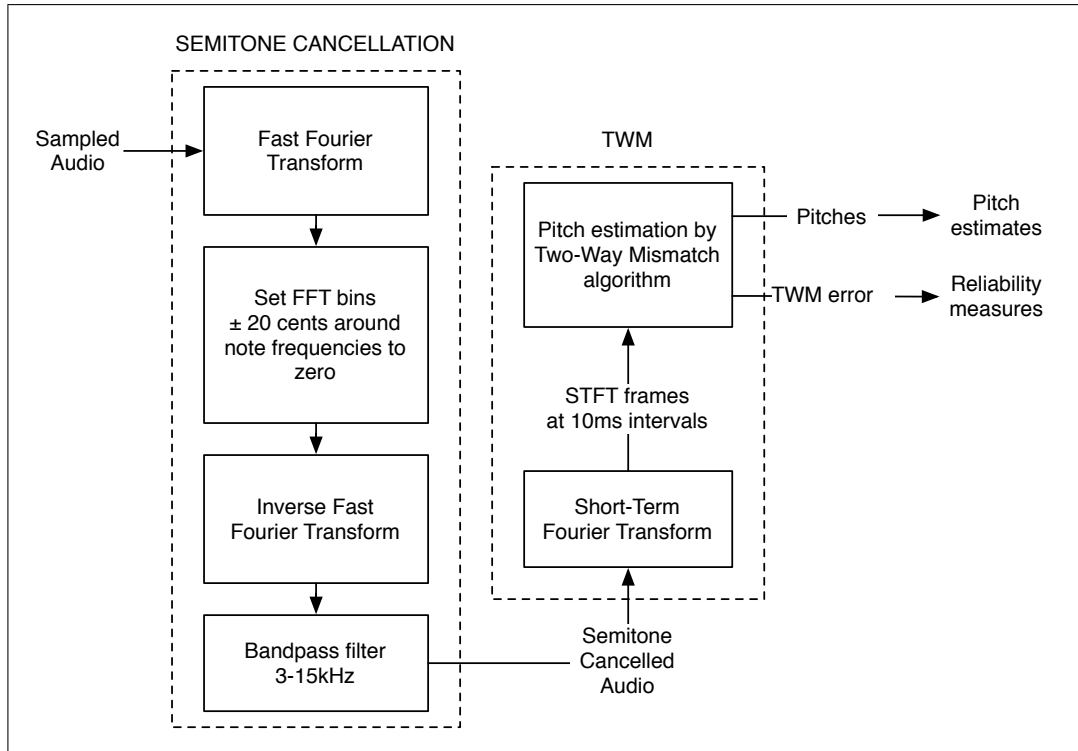


Figure 3.2: Semitone-Cancellation/TWM System Design

ments and then determine which belong to the voice, the whole sound is subjected to a procedure which attenuates any partials not displaying the typical behaviour of the voice. The algorithm they describe, inverse comb filtering, is designed to attenuate all harmonic frequencies of each note in the detected key of the music. Since vocal partials only rarely exist at the exact harmonic frequencies of musical notes, their energy persists and vocal/non-vocal segmentation may be performed by consideration of the energy of the output sound signal.

The inverse comb filtering technique was implemented as part of the initial work for this report, and was found to be overly destructive of vocal melodies. The algorithm was therefore modified in several ways to better suit the task of vocal transcription.

3.2.1 Semitone cancellation algorithm

As mentioned above, attenuating the frequencies of all partials of notes in the detected key was overly destructive. Depending on the degree of attenuation, either many non-vocal notes were audible in the output sound, or vocal notes were also eliminated. The proposed algorithm therefore eliminates components at note frequencies only, which leaves the voice audible while successfully eliminating most non-vocal notes.

Since most partials of non-vocal notes will survive this process, the output was bandpass filtered to preserve the vocal range and exclude particularly low regions (occupied mainly by bass instruments) and particularly high regions (in which non-vocal partials survive the cancellation procedure). This was initially chosen to be

200 Hertz to 2000 Hertz, as specified in [13], but it was found experimentally that 300 Hertz was a more suitable lower limit.

Although informal listening tests suggested most non-vocal notes were successfully attenuated, some remained, most notably in jazz music (where many notes played do not belong to the musical key). The final algorithm therefore performs no key detection, and simply eliminates the fundamental frequencies of all semitone notes.

For the sake of efficiency and flexibility, the Fast Fourier Transform (FFT) was used rather than an inverse comb filterbank (which would have to be carefully re-designed for each modification to the system during development). Since the procedure is time-invariant, a single transform of the whole signal may be performed and operated upon, before performing the inverse transform to obtain the modified sound signal. In place of the notches of the inverse comb filters, the algorithm sets FFT bins in a range around the desired frequency to zero. This range is defined as a number of cents, such that the number of bins affected increases with the target frequency to match the spacing of semitone frequencies. Based on the properties of vocal vibrato compared with other instruments (see Section 2.5) a value in the range 15 to 30 cents seemed appropriate. It was found experimentally that ± 20 cents was a suitable range to eliminate.

Since a single Fourier Transform is used, the frequency resolution achieved will depend on the length of the signal. To achieve 1-cent resolution at the lowest frequency note under consideration (73.42Hz), the FFT length must be just over one million points, corresponding to a signal length of about 24 seconds at 44.1kHz sampling rate. For the 15-second training samples used, a resolution of about 1.5 cents at the lowest note is obtained, which is still sufficient for the semitone cancellation process.

In summary, the semitone cancellation algorithm first performs a Fast Fourier Transform of the entire signal. Then, for each semitone above D2 (73.42Hz), the FFT bins corresponding to ± 20 cents around the semitone frequency are set to zero. The inverse transform is used to produce the semitone-cancelled output sound, which is bandpass filtered to 300–2000Hz.

3.2.2 Pitch estimation by TWM

Informal listening tests verified that the output from the semitone cancellation system is roughly monophonic. Although components of other instruments survive, it is generally unpitched parts such as guitar strumming, piano hammer hits and percussion which are present, and any non-vocal pitched sounds which do survive the cancellation procedure are generally much quieter than the vocal notes.

The sound is therefore passed to a monophonic transcription algorithm which will prioritise sounds with strong partials over background noise and weak spectral peaks. The algorithm chosen is the Two-Way Mismatch algorithm described in Section 2.2.1.

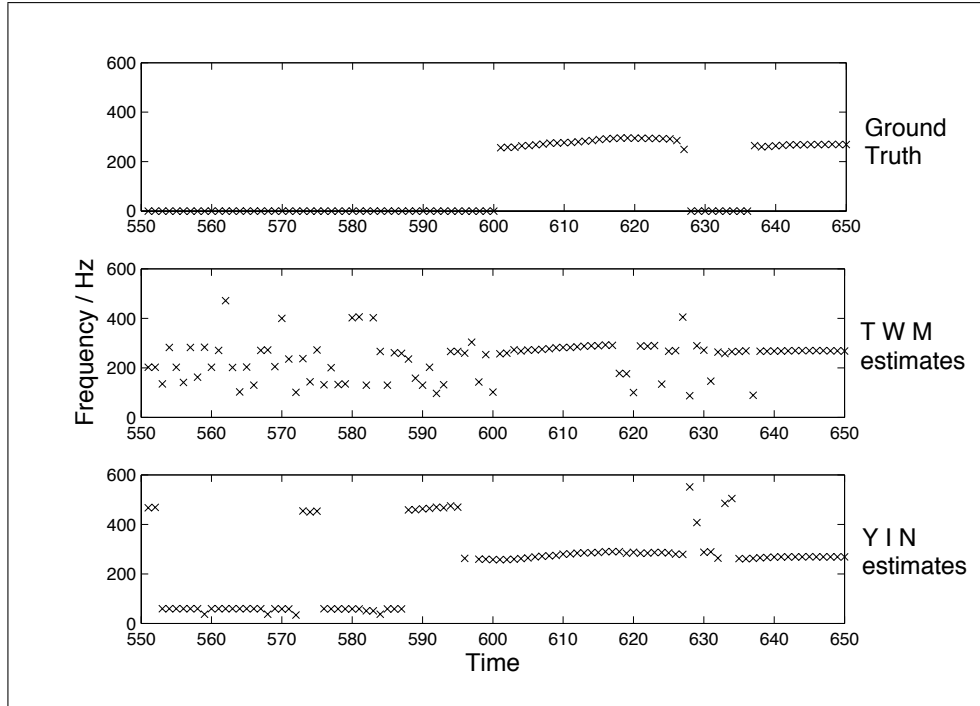


Figure 3.3: An example of the unvoiced frame behaviour of the TWM and YIN transcription systems.

A Matlab implementation of the algorithm, together with suitable Short-Term Fourier Transform front-end is made available with the book “*DAFX : Digital Audio Effects*”[14] as part of an implementation of Sinusoidal Modelling Synthesis. This was used with only minor modifications to set the output time frame spacing to 10ms and produce the reliability measures described below.

One important factor in the choice of pitch estimation algorithm was its output during unvoiced frames. The desired behaviour was that pitch estimates for frames without vocals would be uncorrelated with one another. That is, the estimation system does not reliably transcribe very weak instrumental notes which survive the semitone cancellation process. This allows for good V/UV segmentation by the modified HMM.

An alternative monophonic transcription system is the YIN algorithm[5] which was also tried during development. While it showed comparable accuracy to the TWM algorithm during voiced frames, its behaviour during unvoiced frames was generally to transcribe very weak notes. An illustration of this can be seen in Figure 3.3 which shows the ground truth, TWM estimates and YIN estimates for a one second excerpt from one of the training samples. While different parameters for the YIN algorithm prevented this behaviour somewhat, it was at the cost of accuracy during voiced frames.

The voiced frame accuracy for the two systems on the training samples is shown in Table 3.1. For each sample, a measure of the “incoherence” of estimates during unvoiced frames is also shown. The measure is calculated as follows : for each estimate during an unvoiced region, the pitch difference (in cents) to its successor is recorded. The proportion of such transitions which are greater than the maximum

Training sample :		1	2	3	4	5	6	7	8	9
TWM	Pitch Accuracy (%)	55	35	64	63	68	63	50	74	69
YIN	Pitch Accuracy (%)	56	42	66	60	60	49	37	83	69
TWM	Incoherence	0.74	0.70	0.66	0.58	0.73	0.69	0.70	0.44	0.58
YIN	Incoherence	0.19	0.22	0.34	0.18	0.27	0.30	0.30	0.13	0.21

Table 3.1: Comparison of TWM and YIN for transcribing semitone-cancelled audio.

pitch transition across all training samples (found to be 257 cents) defines the system’s “incoherence” during UV frames. An estimation system which displays the desired behaviour during unvoiced frames should have a high incoherence measure, while a low measure indicates stable pitches are being estimated during vocal silence.

As can be seen from the table, the pitch accuracy during voiced frames was comparable for the two systems, but the TWM system displays considerably more scattering during unvoiced frames.

The correlogram designed below (Section 3.3.2) was also tested as a possible transcription method for the semitone-cancellation system, using all 50 channels. It was found that the ACFs computed from the semitone-cancelled audio very rarely displayed clear peaks, and so the resulting pitch estimates were almost entirely incorrect. Unfortunately, time did not allow for it to be established whether this is due to an incompatibility of the correlogram with the semitone-cancellation procedure, or whether tailoring of the correlogram design to better suit the context could produce reasonable results.

The TWM system was therefore chosen as providing good pitch estimation accuracy and the potential for good V/UV segmentation in post-processing.

3.2.3 Reliability measure

To combine estimates from multiple pitch estimation systems, some method for choosing between estimates is required. If a large number of estimation systems were being used, a majority-rule approach might be suitable, in which the most popular pitch estimate among all systems is chosen. With only two systems this is not possible and some other measure of reliability is required.

In frames when the voice is drowned by percussion, or another instrument’s note has survived the cancellation procedure, or the voice is silent and the sound is mainly background noise and percussion, a pitch will still be estimated but its TWM error will be high. These are the kinds of estimates we wish to mark as “unreliable”, and so the TWM error is a good measure of unreliability. A reliability measure can be easily derived by subtracting the TWM error from a suitable threshold (in this case 1.6) and limiting its range to $[0, 1]$.

The distributions of reliability values for correct and incorrect estimates from this system are calculated as part of the design of the modified HMM (Section 3.4.2),

and verify that the TWM error is a suitable indicator of reliability.

3.3 High-Frequency Correlogram

To investigate the vocal specificity of the transcription system by Li and Wang[11], a series of experiments was carried out to verify a suspected cause (vocal dominance in high frequencies). Based on the results of those experiments, a correlogram pitch estimation system with high-frequency bias was implemented.

The experiments to verify that the voice exhibits greater energy in high frequencies than other instruments are described in Section 3.3.1, along with their results. Next, the design of the system and experiments to further determine a suitable frequency range are presented (Section 3.3.2). A simpler approach to combining channel estimates than that of Li and Wang was used and is described in Section 3.3.3. The reliability measure for this estimation system is given in Section 3.3.4.

3.3.1 High frequency dominance experiments

The paper by Li and Wang demonstrated that their chosen approach to pitch estimation did reliably prefer vocal pitches. However, since the cause of this was unclear, a series of experiments were carried out to test the theory that vocal notes must dominate in high frequencies for the high frequency correlogram channels to reliably prefer the vocal pitch.

The voice was tested against six classes of instrument :

- Keyboard (inc. traditional piano, electric piano, synthesised organ)
- Plucked Strings (inc. acoustic, electric, electric slide and 12-string guitar, banjo, mandolin)
- Bowed Strings (inc. solo violin, orchestral string section)
- Wind (inc. harmonica, solo saxophone, brass and woodwind sections)
- Bass (inc. upright, electric and synthesised bass)
- Percussion (inc. rock drum sets in a variety of styles, synthesised drum machine)

For each class, ten recordings were gathered covering the range of instruments shown above and resulting in a total of 40–180 seconds of audio per class. The ten vocal recordings consisted of five male and five female extracts.

For each recording, a Short-Term Fourier Transform was performed (with 50% overlapping 10ms windows) and converted into a measure of per-band power. This was summed over frames to provide a measure of per-band energy, which was normalised to sum to unity. Finally this was converted into a decibel measure for

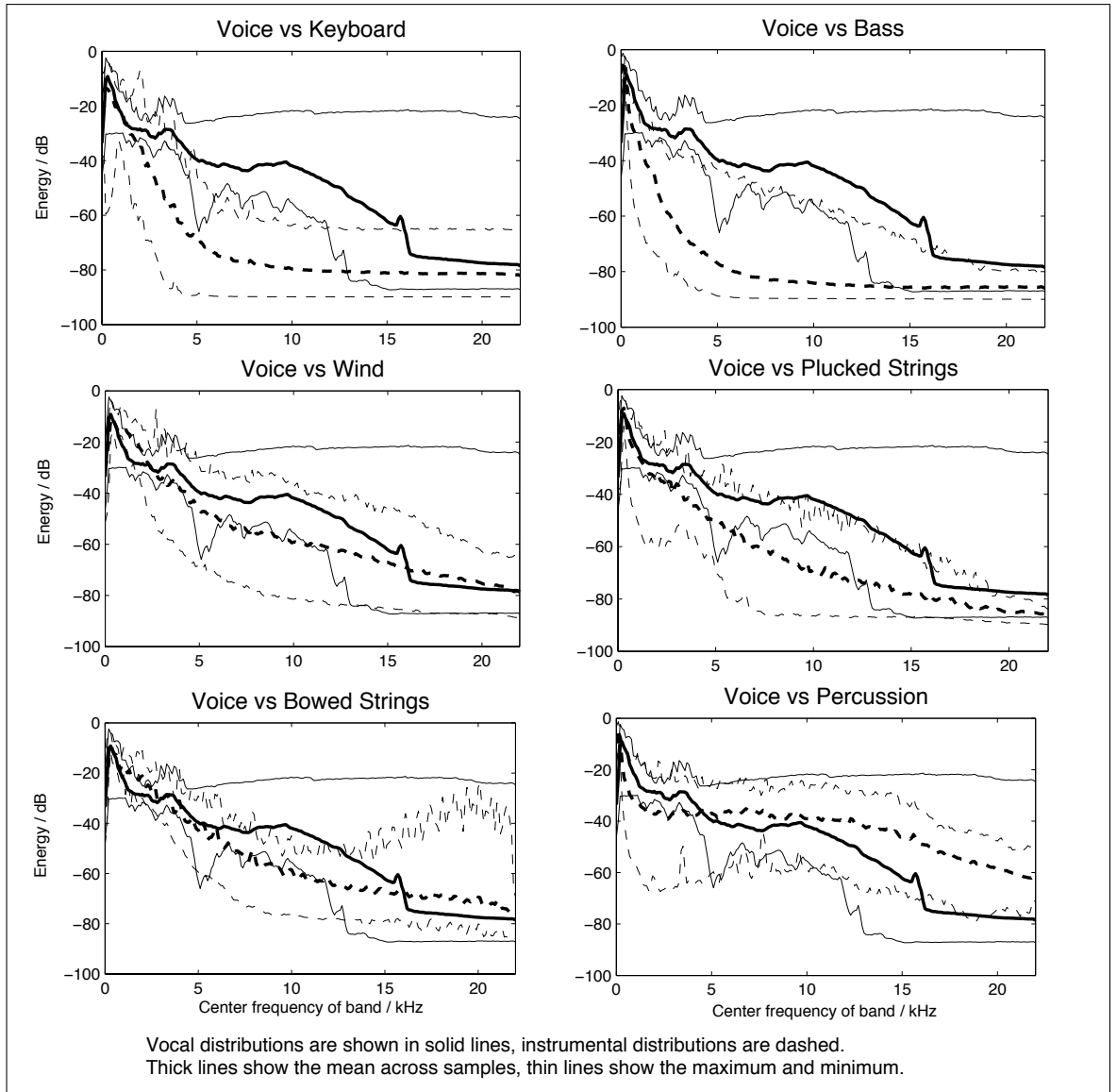


Figure 3.4: Energy distributions of human singing voice compared with other instruments.

that recording. For each class, the minimum, mean and maximum values for each band across recordings was calculated to provide an overall indication of that class's minimum, average and maximum decibel energy in each frequency band.

Graphs comparing the voice's per-band loudness with that of each instrument class are shown in Figure 3.4. In the case of Keyboard and Bass, the voice can be seen to quite reliably dominate in the 3–12kHz range, with the minimum vocal loudness roughly coinciding with the maximum instrumental loudness. The Wind and Plucked Strings cases are less clear-cut, but the average vocal loudness is still considerably higher (7–30dB) than the average instrumental loudness in the 3–15kHz range.

The case of Bowed Strings is less promising, with a higher minimum instrumental loudness, and a reduced frequency range of around 6–15kHz for the average vocal loudness being greater. There is an additional concern since bowed strings are the

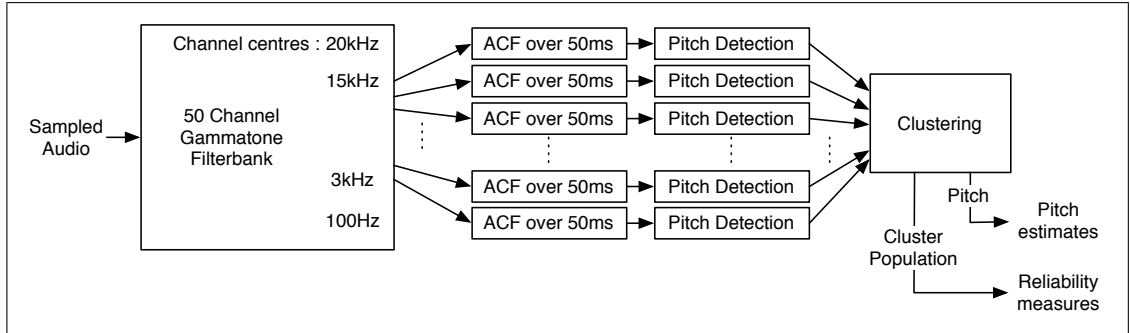


Figure 3.5: High-Frequency Correlogram System Design

class most likely to exhibit considerable vibrato which would cause difficulties in the other pitch estimation system.

Finally, it can be seen that percussive loudness is quite evenly spread across frequencies, with the vocal average dominating only in a small, low frequency range. Fortunately (for the sake of music as well as these experiments !) percussive sounds are very short-lived compared with instrumental sounds, and their detrimental effect on pitch estimation is inevitable but localised.

These experiments confirm the suspected high-frequency dominance of the human singing voice, and suggest a frequency range of 5–12kHz for estimating vocal pitches. The next section includes details of further experiments to establish a reliable frequency range once the estimation system was built.

3.3.2 Design of high-frequency correlogram

The design of the high-frequency correlogram was based on that used by Li and Wang, but the final system differs quite substantially. An overview of the system is shown in Figure 3.5. Since experiments had indicated that higher frequency channels than 8kHz might be useful and downsampling to 16kHz would make them unavailable, the signal was kept at 44.1kHz. Where Li and Wang used 128 channels in the range below 8kHz, it was found that 50 channels in the range 100Hz–22kHz was sufficient. As is explained below, not all of these channels are actually used, but the gammatone filterbank design is based on these parameters. The Auditory Toolbox for Matlab[38] was used to design and apply the filterbank.

The system by Li and Wang estimated pitch from ACFs calculated on 16ms frames spaced by 10ms. It was found during development that using such short frames often produced ACFs without reliable peaks for pitch estimation, perhaps due to the high frequency information having been retained. A longer frame length of 50ms was therefore chosen. The unbiased ACF calculation specified in Equation 2.2 is used to reduce estimation error at higher lags. The resulting ACF is normalised by dividing by the zero-lag value and subtracting the mean.

The combing approach to pitch estimation from ACFs[15] was found to be biased towards low-frequency (large lag) pitches where only a few multiples of the base lag

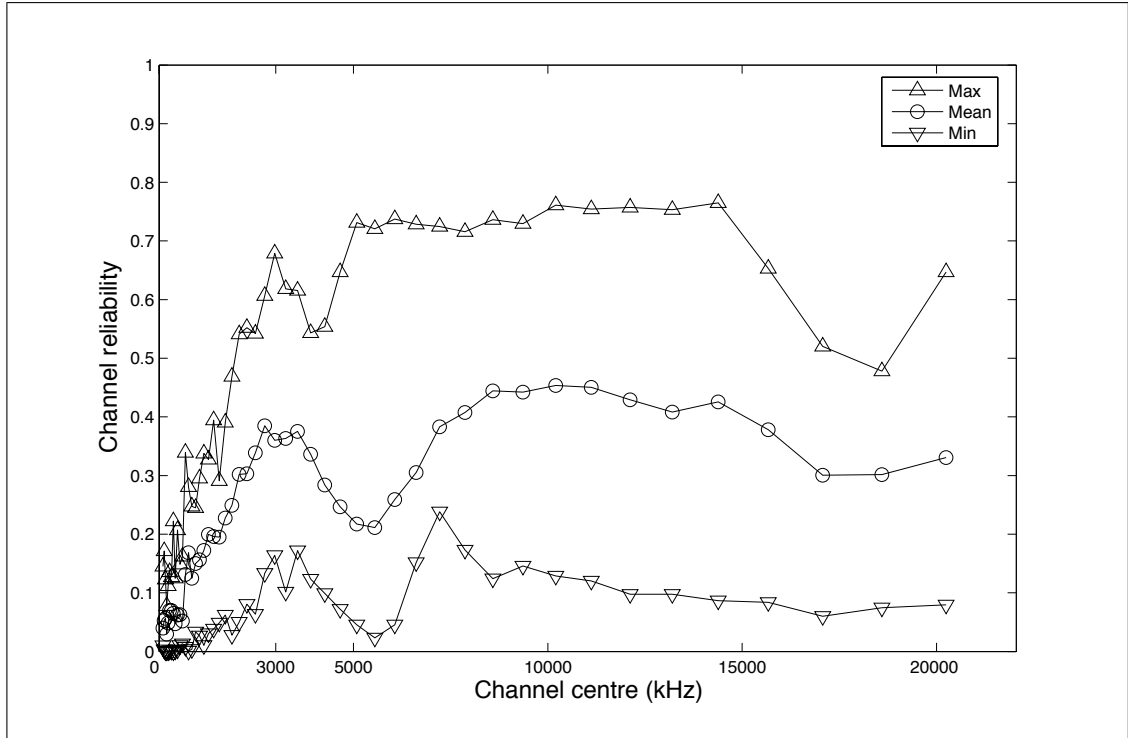


Figure 3.6: Reliability of correlogram channels for training data.

exist in the ACF range under consideration. The previous application of this approach, beat tracking, used a perceptual model for tempos to weight lag candidates and avoid such behaviour. There was not an obvious equivalent for pitches in the vocal range, but it was found that considering only the first three lag multiples for all pitch candidates produced good pitch estimates.

The pitch is therefore estimated from this normalised ACF by considering each pitch in the singing voice range (80–1000Hz) which corresponds to an integer lag value and summing the ACF value at the first three multiples of this lag to produce a score value for that pitch. The pitch with greatest score is selected.

The system described above produces a pitch estimate per channel at 10ms intervals. This allowed for further investigation of the theory of high frequency vocal dominance and more careful selection of a suitable channel range. Since the spectral presence of an instrument can be greatly altered by the sound engineer in mixing a track, it is important to measure the usefulness of channels across a range of recordings. For each of the nine training samples, the reliability of each of the 50 channels was measured as the proportion of estimates correct (within 50 cents of ground truth) in voiced frames. The minimum, mean and maximum of channel reliabilities across training samples was then calculated, and can be seen in Figure 3.6.

The mean reliability plot shows that on average, channels above 3kHz are more reliable estimators of vocal pitch than those below 3kHz. There is a clear dip in reliability between 4kHz and 7kHz, most likely caused by filtering during the mixing of the track. In determining the full range of channels which may prove useful for vocal pitch estimation, the maximum reliability plot is the most relevant as it

indicates the potential of each channel to be useful in transcribing a new recording. Ignoring the slight dip mentioned above, this plot shows high values for the range 3–15kHz.

Based on these measurements of channel reliability, it was decided that channels in the range 3–15kHz were reliable estimators of vocal pitch. The system therefore computes the gammatone filterbank output only for the 19 channels whose centre frequencies lie in this range. This reduces computation time and should prevent non-vocal pitches from being transcribed.

3.3.3 Combination of channel estimates

The high-frequency correlogram as described above produces 19 pitch estimates per time frame and so some method for combining them to produce a single estimate was required. In the paper by Li and Wang, a statistical approach is taken to combine low channel information with high channel information. Since we are computing estimates only for those channels which are assumed to reliably estimate vocal pitch, a simpler approach may be taken, in which channel estimates are clustered by frequency and the cluster with highest population chosen as the output estimate.

Initially an algorithm based on k-means clustering was implemented, which grouped the data into several clusters and combined clusters whose centres were within 50 cents of each other. This was found to permit clusters whose members were too widely spaced, and hence had badly-estimated centres. There is a trade-off between reliably choosing the right cluster and having precisely-calculated cluster centres, and a better compromise was found by using a simple clustering approach in which all neighbouring estimates spaced by less than 50 cents are clustered together.

This approach to combining channel estimates provides the desired behaviour of scattered estimates during unvoiced regions. During frames when the vocals are not present, the channel estimates are less clearly clustered around any single frequency. Weak instrumental notes gain estimates from only a small number of channels, and random estimates caused by noise influencing a channel gather similar populations, such that from frame-to-frame in unvoiced regions, the particular cluster chosen will vary and the estimates will be scattered from frame to frame.

The pitch accuracy and “incoherence” scores for this system are shown in Table 3.2. For the sake of comparison, a High-Frequency TWM system was implemented by modifying the TWM algorithm to only measure or prototype peaks in the 3–15kHz range. The scores for this system are also shown in the table.

As can be seen in the table, the correlogram approach outperforms the TWM approach considerably in terms of pitch accuracy. The incoherence scores are similar between the two systems, but given the low pitch accuracy of the TWM approach, this does not necessarily indicate different behaviour during unvoiced periods — calculating the TWM system’s incoherence scores for voiced frames confirmed that its estimates are in fact scattered just as much during voiced regions.

Training sample :		1	2	3	4	5	6	7	8	9
HF-Corr.	Pitch Acc. (%)	64	68	66	86	74	34	32	58	69
HF-TWM	Pitch Acc. (%)	25	10	23	57	29	32	16	13	31
HF-Corr.	Incoherence	0.59	0.58	0.50	0.50	0.71	0.41	0.49	0.52	0.44
HF-TWM	Incoherence	0.58	0.67	0.59	0.57	0.63	0.44	0.19	0.29	0.46

Table 3.2: Comparison of Correlogram and TWM systems for transcribing with High-Frequency bias.

The YIN algorithm was also tried, first on audio bandpass-filtered to the 3–15kHz range, then as a per-channel pitch estimation system after the gammatone filterbank was applied. Neither approach achieved even low estimation accuracies. As with the other systems implemented for comparison’s sake, it is possible that low accuracy scores are a result of a lack of time spent adjusting the system. However, it is also possible that the systems are simply not appropriate for the task. Determining which is the case would require considerable time which was instead spent pursuing the primary systems (TWM for semitone-cancellation and the correlogram for high-frequency bias).

3.3.4 Reliability measure

As with the semitone-cancellation/TWM pitch estimation system, some measure of the reliability of estimates is required. In this case, we wish to mark estimates as reliable when there is a strong consensus among channels, and unreliable when there is not a single dominant cluster. The proportion of channel estimates belonging to the chosen cluster therefore serves as a good measure of reliability. This is demonstrated in Section 3.4.2 where the distributions of reliabilities for correct and incorrect estimates from the high-frequency correlogram system are computed.

3.4 Modified Hidden Markov Model

This section details the design and implementation of the modified Hidden Markov Model used to combine pitch estimates from the two systems described above and perform V/UV segmentation to produce a final transcription of the vocal melody. The use of HMMs to perform post-processing on pitch estimates is quite well established, but for a number of reasons a novel approach was required.

Firstly, the input consists of pitch estimates from multiple systems, with associated reliability measures. Observation probabilities in the HMM would have to be designed for this if using multiple pitch estimation systems were to prove advantageous, and this is discussed in Section 3.4.2, below.

Secondly, since vocal melodies are so far removed from the idealised piano-roll type score, the desired output transcription should capture the pitch fluctuations of the voice, and hence not perform quantisation to semitone frequencies. The note modelling HMM by Ryyänen and Klapuri introduced in Section 2.3 is designed to

have hidden states corresponding to the singer’s intended notes, and observations corresponding to the actual pitches produced. The HMM proposed here instead has hidden states corresponding to the actual pitch sung, and observations corresponding to the estimated pitches from the two estimation systems. This means that the set of hidden states must reflect a continuous range of frequencies, an issue which is tackled in Sections 3.4.1 and 3.4.4.

Thirdly, the novel method of V/UV detection proposed here (of inferring silence from scattered pitch estimates and low reliability values) requires implementation in the HMM, by means of the observation probabilities (Section 3.4.2) and also the transition probabilities which are described in Section 3.4.3.

After the model and the learning of its parameters have been described, some necessary modifications to the Viterbi[12] algorithm for inferring the most probable state path are discussed in Section 3.4.4.

3.4.1 Dynamic state generation for continuous pitch

The classical HMM is defined by a set of states, observation probabilities and transition probabilities, which remain fixed over time. To permit a continuous range of frequencies in a set of hidden states would require an infinite number of states, which is naturally impractical. The solution proposed here can be seen in two ways. First, as allowing the set of states defining the HMM to be dynamic with respect to time, or alternatively as defining a set of infinitely many hidden states, only a subset of which may be occupied in each time frame. The discussion here takes the former viewpoint, but could equally well be presented from the latter.

The first step towards the dynamic state generation algorithm is to let the hidden states of the model be defined at each time instant by the pitch estimates corresponding to that instant. For a system with K pitch estimators, the HMM will have K hidden states at each time instant, whose nominal frequencies are defined by the pitch estimates for that instant. This simple system would serve to allow selection of one of the K estimates for each instant in time. For the system proposed in this report there are two pitch estimators and so $K = 2$.

Additionally (to allow V/UV segmentation) let the state corresponding to $j = 0$ represent silence, such that for K pitch estimates $e_{k,t}$, $k = 1 \dots K$ at time t , the set of $(K + 1)$ states of the HMM at time t is defined by :

$$\Omega_t = \{\omega_{j,t}\}_{0 \leq j \leq K} \quad (3.1)$$

where

$$\omega_{j,t} = \begin{cases} \text{silence,} & \text{if } j = 0 \\ e_{j,t}, & \text{if } j = 1 \dots K \end{cases} \quad (3.2)$$

In what follows, the notation $\omega_{j,t}$ is used to refer both to states and their assigned frequency values. Similarly, $e_{j,t}$ is used to refer to pitch estimates and also the values they take on.

Since pitch estimation systems can produce incorrect estimates during voiced regions, and we cannot assume any of the other systems are providing the correct pitch, there may not be a state corresponding to the true pitch, and the HMM would be forced to occupy the silent state or an incorrect pitch state. To avoid this, the system generates a *dummy state* at time t for each state at time $t-1$ for which there is no estimate at time t within 50 cents of its frequency. The frequency value of the dummy state is that of the state at $t-1$ which generated it. Mathematically, $\omega_{j,t-1}$ is added to the set of states Ω_t if and only if there is no k for which $e_{k,t}$ is within 50 cents of $\omega_{j,t-1}$:

$$(\nexists k \text{ s.t. } (1200 \times \left| \log_2 \left(\frac{\omega_{j,t-1}}{e_{k,t}} \right) \right| \leq 50)) \Rightarrow \omega_{j,t-1} \in \Omega_t \quad 1 \leq j \leq |\Omega_{t-1}| \quad (3.3)$$

where $(1200 \times \left| \log_2 \left(\frac{f_1}{f_2} \right) \right|)$ gives the difference in cents between two frequencies f_1 and f_2 .

This condition would lead to a steadily increasing number of states, most of which would be unsupported by any recent estimates. A *pruning* process is therefore introduced, to periodically remove dummy states which have not been recently used in any of the most probable paths ending in the current set of states. This is discussed further in Section 3.4.4 below.

3.4.2 Observation probabilities based on all observed estimates

Unlike previous post-processing HMM systems, which use a single pitch estimate to compute the observation probability for each state, the system proposed here considers all estimates for each state. That is, the likelihood of the HMM occupying a state depends not only on the pitch estimate closest to that state's frequency (as in the system[3] described in Section 2.3), but on all estimates from all pitch estimation systems. This approach, combined with the use of reliability measures on estimates permits reliable choice of pitch estimate, and aids V/UV segmentation.

The observation probability for a given state is calculated as the product of its observation probabilities for each estimate :

$$P(\{e_{k,t}\}_{1 \leq k \leq K}, \{r_{k,t}\}_{1 \leq k \leq K} \mid \omega_{j,t}) = \prod_{k=1 \dots K} P(e_{k,t}, r_{k,t} \mid \omega_{j,t}) \quad (3.4)$$

where $r_{k,t}$ is the reliability of the k 'th estimate at time t , $e_{k,t}$, and $\omega_{j,t}$ is the state whose observation probability is being computed.

This combination of observation probabilities by multiplication is based on an assumption that the observations for different pitch estimation systems are independent. While this is unlikely to truly be the case, a large amount of training data would be required to learn a combined probability distribution for multiple estimation systems, and the product distribution should prove to be a reasonable approximation to the true distribution.

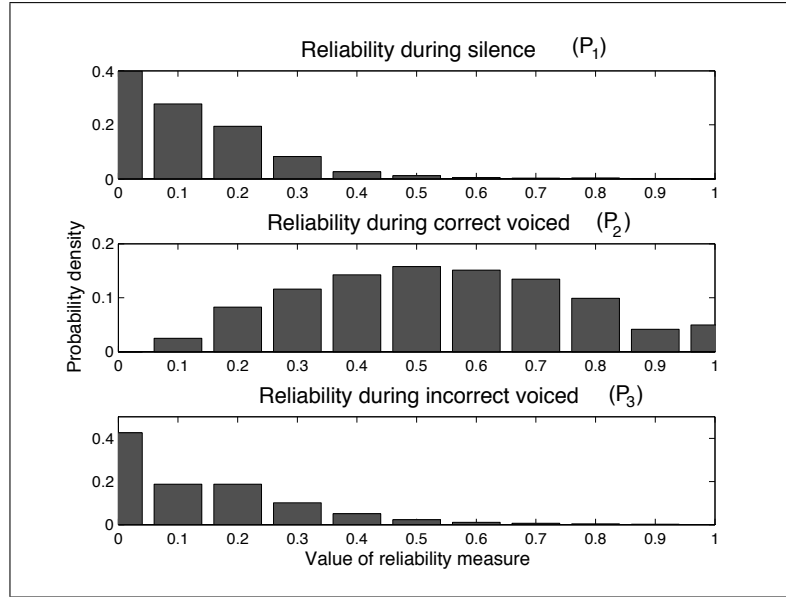


Figure 3.7: Distributions of reliability measures from the High-Frequency Correlogram system

The observation probabilities for each estimate are calculated using three probability distributions which are specific to the pitch estimation system which produced the estimate :

$$P(e_{k,t}, r_{k,t} | \omega_{j,t}) = \begin{cases} P_{1,k}(r_{k,t}) & \text{if } j = 0 \\ P_{2,k}(r_{k,t}) & \text{if } j \neq 0 \text{ and } (1200 \times \left| \log_2 \left(\frac{\omega_{j,t}}{e_{k,t}} \right) \right| \leq 50) \\ P_{3,k}(r_{k,t}) & \text{if } j \neq 0 \text{ and } (1200 \times \left| \log_2 \left(\frac{\omega_{j,t}}{e_{k,t}} \right) \right| > 50) \end{cases} \quad (3.5)$$

That is, the observation probability, $P(e_{k,t}, r_{k,t} | \omega_{j,t})$, for state j and estimate k is defined using a single distribution parameterised by the reliability measure ($r_{k,t}$), with the choice of distribution P_1 , P_2 or P_3 depending on whether the state is silence or a voiced state and whether the estimate frequency is within 50 cents of the state frequency. The condition of the estimate frequency being within 50 cents of the state frequency is based on the evaluation criterion (introduced in Section 2.4.4) which marks pitch estimates as “correct” if they are within 50 cents of the ground truth.

Approximations to the distributions P_1 , P_2 and P_3 were calculated by applying the two pitch estimation systems to the nine training recordings mentioned at the start of this chapter. Distributions of the reliability measures were gathered for each of the two pitch estimation systems, based on whether the ground truth was silent or voiced, and in the latter case whether the pitch estimate was within 50 cents of the ground truth pitch. Histograms of the gathered values then serve as models of the probability distributions. The estimated distributions are shown in Figures 3.7 and 3.8, calculated for bins centered on $\{0, 0.1, 0.2 \dots 1\}$.

For both systems it can be seen that the reliability measures during silence, and for incorrect pitch estimates are generally low, while those for correct pitch

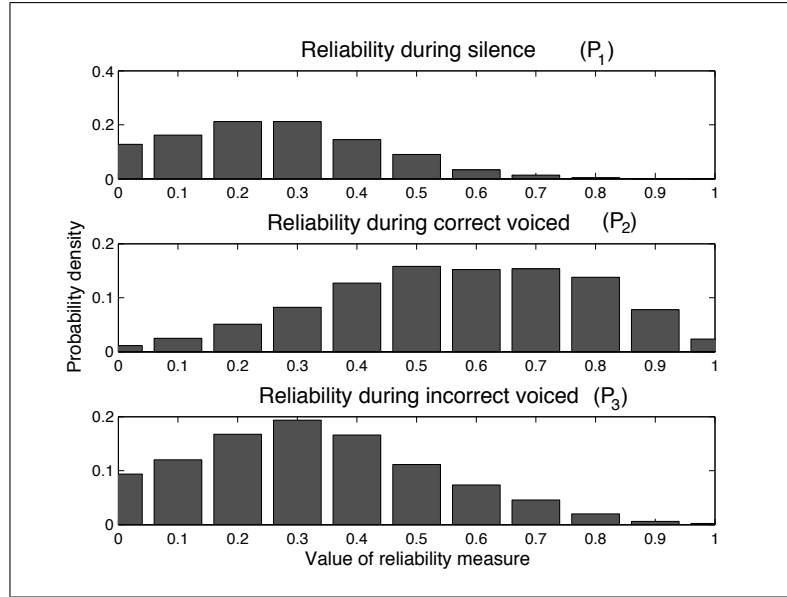


Figure 3.8: Distributions of reliability measures from the Semitone-Cancellation/TWM system

estimates are high. There is not a clear separation between the distributions such that a threshold could reliably distinguish the cases, but there is sufficient separation that certain values strongly indicate whether the estimate is correct or not.

With these distributions in mind, some examples of the observation probability calculation may be given. In what follows it is not important which estimation system is numbered $k = 1$ and which is $k = 2$, as the distributions have similar shapes.

First, consider the case where only one estimate has high reliability, for example $r_{1,t} = 0.7$, $r_{2,t} = 0.1$. The observation probability of the state $\omega_{1,t}$ corresponding to $e_{1,t}$ will be calculated as the product of two high values ($P_{2,1}(0.7)$ and $P_{3,2}(0.1)$), while the state $\omega_{2,t}$ corresponding to $e_{2,t}$ will have observation probability calculated as the product of two low values ($P_{3,1}(0.7)$ and $P_{2,2}(0.1)$). The observation probability for the silent state would be calculated as the product of a high value ($P_{1,2}(0.1)$) and a very low value ($P_{1,1}(0.7)$). The state corresponding to the high-reliability estimate will therefore have higher observation probability than other voiced states or the silence state.

Another example would be when both estimates have high reliability. The observation probabilities for the corresponding states would be calculated as the product of a high probability and a low probability, while the observation probability for the silent state would be the product of two low probability values. The voiced states would be considered similarly likely on the basis of the observations, while the silent state would be less likely than either.

Finally, when both estimates have low reliabilities, their corresponding states would have observation probabilities which are the product of a low and high value, while the silent state's observation probability would be the product of two high

values. The silent state would therefore tend to have higher observation probability than any voiced state.

Based on observation probabilities alone, we can see that the system will tend to prefer states with supporting high-reliability estimates when other estimates have low reliability. It will consider such states approximately equally when there are multiple high-reliability estimates. It will prefer silence when all estimates have low reliability.

3.4.3 Transition probabilities and V/UV segmentation

The transition probabilities for the HMM are designed to take advantage of the unvoiced region behaviour of the two pitch estimation systems, by restricting the pitch difference between successive voiced states. This means that when pitch estimates become scattered, dummy states or silence will be chosen.

The transition probability from state i to state j is defined as :

$$P(\omega_{j,t} | \omega_{i,t-1}) = \begin{cases} P(\textit{silence} \rightarrow \textit{silence}), & i = 0, j = 0 \\ P(\textit{silence} \rightarrow \textit{voiced}) \times \frac{1}{|\Omega_t|-1}, & i = 0, j \neq 0 \\ P(\textit{voiced} \rightarrow \textit{silence}), & i \neq 0, j = 0 \\ c_{i,t} \times (P(\textit{same_note}) \times e^{-\frac{d^2}{100}} + P(\textit{change_note}) \times e^{-\frac{d^2}{200}}), & i \neq 0, j \neq 0 \end{cases} \quad (3.6)$$

where d measures the pitch difference in cents :

$$d = 1200 \times \left| \log_2 \left(\frac{\omega_{i,t-1}}{\omega_{j,t}} \right) \right| \quad (3.7)$$

and $c_{i,t}$ is a normalisation factor, chosen such that the transition probabilities from a particular state will sum to one.

The fourth case is a model of the pitch change for transitions between voiced states. It is a simple combination of Gaussian models for the variation in pitch during a single note, and in the transition between notes. The two Gaussian distributions have variances of 50 and 100 cents respectively, and are normalised to have unit value for a zero input, resulting in the simple exponential expressions of Equation 3.6. The values for the two variances were chosen based on consideration of the ground truth pitch contours for the training data. It was observed that strong vibrato could lead to considerable frame-to-frame changes in pitch, in the order of 50 cents or so, while transitions between notes tended to be quite smooth such that the frame-to-frame differences were rarely greater than 100 cents or so. The resulting overall model for pitch changes between voiced states is shown in Figure 3.9.

The actual estimated probability distribution of frequency differences between successive voiced frames in the training data was tried in place of this simple model, but found to be a much narrower distribution which resulted in less accurate transcriptions. This is thought to be due to a lack of data resulting in a poor estimation of the true distribution, which is perhaps better approximated by the simple model described above.

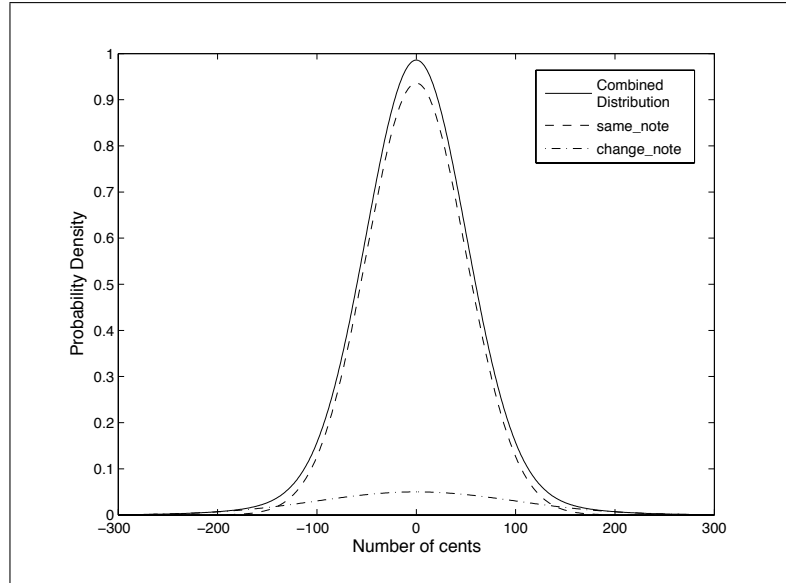


Figure 3.9: Model distribution of pitch changes between successive voiced states, along with its two component distributions.

The probability constants in Equation 3.6 above were estimated from the training data as :

$$P(\textit{silence} \rightarrow \textit{silence}) \approx 0.97 \quad (3.8)$$

$$P(\textit{silence} \rightarrow \textit{voiced}) \approx 0.03 \quad (3.9)$$

$$P(\textit{voiced} \rightarrow \textit{silence}) \approx 0.014 \quad (3.10)$$

$$P(\textit{same_note}) \approx 0.936 \quad (3.11)$$

$$P(\textit{change_note}) \approx 0.05 \quad (3.12)$$

The first three cases are measured directly from the ground truth for the nine training recordings, while the last two are derived from the measured value of $P(\textit{voiced} \rightarrow \textit{voiced}) \approx 0.986$ based on approximate relative frequencies of “same note” and “change note” transitions.

3.4.4 Viterbi algorithm for state sequence inference

Once the model has been designed and trained, a method for inferring the most likely state sequence given a set of observations is required. The Viterbi algorithm[12] was chosen, with some modifications being necessary to suit the modified HMM described above.

The main modification to the algorithm is the “pruning” procedure, mentioned in Section 3.4.1 above, which prevents the number of states from increasing without bound and avoids the presence of dummy states which are unsupported by any recent estimates.

This is a straightforward addition to the Viterbi algorithm, implemented as a short routine at the start of each time frame’s processing. The pruning process can be described by modifying the condition of Equation 3.3. For states corresponding

to estimates at $t - 1$ the condition remains unchanged, but dummy states at $t - 1$ are perpetuated only if they were used in the last two time frames by one of the most probable paths calculated by the Viterbi algorithm :

$$(\nexists k \text{ s.t. } (1200 \times \left| \log_2 \left(\frac{\omega_{j,t-1}}{e_{k,t}} \right) \right| < 50)) \Rightarrow \omega_{j,t-1} \in \Omega_t \quad 1 \leq j \leq K \quad (3.13)$$

$$(\nexists k \text{ s.t. } (1200 \times \left| \log_2 \left(\frac{\omega_{j,t-1}}{e_{k,t}} \right) \right| < 50)) \wedge (\exists i, c \text{ s.t. } p_{i,t-c} = \omega_{j,t-c-1}) \Rightarrow \omega_{j,t-1} \in \Omega_t \\ K < j \leq |\Omega_{t-1}|, \quad 1 \leq c \leq 2 \quad (3.14)$$

where $p_{j,t}$ are the backpointers for the Viterbi algorithm, such that $p_{i,t} = j$ implies that state j is the most probable $(t - 1)$ state for a state sequence ending in state i at time t .

Using this modified Viterbi algorithm to infer the most likely state sequence given sets of estimates from the two pitch estimation systems was found to perform V/UV segmentation quite reliably on the training data, but occasionally very short voiced or unvoiced regions would be transcribed. To avoid this behaviour and model region lengths more realistically, a minimum length of 5 time frames (corresponding to 50ms) for voiced and unvoiced periods was imposed as a modification to the Viterbi algorithm. This means that the model is not a pure Markov model, but rather a simple example of a Hidden Semi-Markov Model[39] in which the state occupancy durations are explicitly modelled. The Viterbi algorithm will find the most likely state path given this added constraint.

4. EVALUATION

This chapter begins by describing the data set assembled to evaluate the proposed system in Section 4.1. Next several sets of tests are introduced, and their results presented and discussed (Section 4.2). The performance of the proposed system is also compared with some existing systems. The chapter ends with a discussion of possible future work in Section 4.3.

4.1 Test Set

Having used the 2005 MIREX training data to train system parameters and perform preliminary testing during the design process, a different set of recordings was required for evaluating the system’s performance. Although the sole input to the system is CD-quality recordings of full arrangements of popular music, a ground-truth transcription would also be necessary in order to assess the system’s output.

In evaluating systems which quantise their transcription to semitone frequencies, hand-annotation of the melody pitch is a possibility. However, since we require the exact frequency of the vocal pitch in each 10ms frame, hand-annotation is infeasible. Instead, if the vocal part is available as a separate recording, a monophonic transcription system can be applied, and its output hand-corrected to avoid octave errors and bad voicing decisions.

Since commercial music is almost invariably available only as a pre-mixed stereo recording, obtaining a separated vocal part is not generally possible. Two solutions to this problem were used in assembling the test set: karaoke CDs and multi-track recordings.

Some older CDs produced for karaoke performances are recorded with the accompaniment in one stereo channel, and an example vocal performance with accompaniment in the other stereo channel. In some cases, the accompaniment is almost identical in both channels (except for a level reduction in one channel to allow for the vocals), and the task of extracting a solo vocal recording can be viewed as a noise reduction problem. The chosen approach to solving this problem is discussed in Section 4.1.1 below.

Another possibility is to obtain multi-track recording data (in which the recording of each instrument is typically available as a separate channel) and produce recordings of the full mix, and of the vocal part alone. Some artists (both professionals and amateurs) choose to make their multi-track data available online to encourage fan “remixes” of their work. The data sourced in this way is described in Section 4.1.2.

Once the solo vocal part has been obtained, some automatic monophonic transcription method may be applied to derive the ground truth melody transcription.

In this case, the speech analysis program “Praat” [40] was used to transcribe the pitch, and this process is discussed in Section 4.1.3 below. Finally a full mix version of the recording must be produced to use as input to the system. The chosen approach is presented in Section 4.1.4.

Nineteen recordings of approximately 30 seconds each were produced, along with corresponding ground-truth text files containing the pitch of the lead vocal part at 10ms intervals (with 0 Hertz indicating silence). The recordings cover a wide range of genres including “golden oldies”, folk, easy listening, jazz, disco, pop, and rock. In contrast with some karaoke CDs, those chosen for the test set have varied and realistic arrangements suited to the musical style of each track.

4.1.1 Wiener filtering on Karaoke CDs

There are a variety of data formats which have been used for karaoke music over the years, and although most modern karaoke CDs contain no “guide” vocal part, some older CDs do. These CDs are referred to as “multiplex CD+G” discs (the “+G” format indicates that the lyrics are available as graphics on the disc), and come in two types. The first type contains each song twice on the CD, once with guide vocals, and once without. The second type contains each song once, using one stereo channel for the accompaniment alone (referred to below as “channel A”), and the other for the accompaniment with guide vocals (“channel B”).

Even within the subset of karaoke CDs recorded in a multiplex format, only some allow the vocal part to be easily extracted. For a given CD to be usable for this purpose, the same accompaniment recording must have been used in both channels/tracks rather than the accompaniment being mixed differently for each channel/track. It was found in assembling the test set that the second type of multiplex CD (which uses a single track per song) tended to meet this requirement.

If the same accompaniment recording has been used for both channels, channel A serves as a good estimator for the “noise” in channel B, and the standard signal processing technique of Wiener filtering [41] can be applied to reduce the noise (ie. remove the accompaniment) in channel B, leaving essentially a pure vocal track.

Seventeen tracks from two CDs were used for the test set, selected so as to avoid tracks with excessive backing vocals. The two stereo channels for each track were extracted, and the first five seconds of audio used to design a tenth order Wiener denoising filter. The filter was applied to channel A, and the result subtracted from channel B to produce the solo vocal recording. Channel A and the resulting vocal recording were saved, with the vocal recording being used later to transcribe the ground truth melody, and both being used to produce the “full mix” recording used as input to the system.

Informal listening tests verified that the accompaniment was barely perceptible in the resulting solo vocal recordings, and the attenuation was certainly sufficient to allow correct pitch estimation as described below.

4.1.2 Multi-track data

It is increasingly common for the internet to be used to allow long-distance musical collaboration, with sites such as MyVirtualBand.com providing the facilities for a physically dispersed group of people to work together on a single track, and amateur and professional musicians alike providing their multi-track data for others to remix.

The Blind Audio Source Separation Database (BASS-dB)[42] collects together multi-track data from a number of amateur musicians, mostly available under a Creative Commons[43] license. In the professional realm, Peter Gabriel recently ran a remix competition in which the individual tracks from one of his songs were made available[44]. Unfortunately no session file was included to reproduce the original mix of those tracks.

Another example of a professional band making their data available is Nine Inch Nails, who have recently made the multi-track session files for two of their songs available online[45] in GarageBand[46] format to encourage fan remixes. Once the session is loaded into the GarageBand program, it is straightforward to disable the lead vocal channel, or all other channels, and hence produce recordings containing just the accompaniment or vocals.

Of all these sources of multi-track data, the Nine Inch Nails tracks proved the easiest to produce vocal, accompaniment and “full mix” recordings from, and were included in the test set along with the karaoke CD data.

4.1.3 Transcription of ground-truth melody

For each recording a 30-second excerpt was chosen, with the main requirement being that any backing vocals present were significantly quieter than the lead vocal part. This excerpt was then analysed to produce the ground-truth melody transcription using the speech analysis program “Praat”.

Although designed for annotating speech rather than music, Praat includes an autocorrelation-based pitch estimation algorithm and a graphical interface for correcting the resulting pitch estimates. The corrections made to pitch estimates for the test set consisted of fixing octave errors (in which the estimates are an octave above or below the true pitch) and incorrect voicing decisions (pitch estimates during silence or consonant noise, or silence transcribed during a note). The “correct” boundaries between voiced and unvoiced regions were judged manually, using a wave editor program to replay particular sections of the recordings.

4.1.4 Combination of vocal and accompaniment recordings

It was originally intended that the input to the proposed system should be the “accompaniment + vocals” channel in the case of karaoke data or the original mix in the case of multi-track recordings. However it was found that in the case of the karaoke data this led to some tracks having very faint or very loud vocal parts

relative to the accompaniment. The approach used by Li and Wang[11], of mixing the solo vocal part with the solo accompaniment part to achieve a 0dB loudness ratio, was found to produce more consistent mixes in which the vocals were clearly audible but not always the loudest instrument playing.

The “loudness” of the accompaniment and vocal recordings is defined as the median of the energies of those (10ms) frames in which the ground truth melody is voiced. The difference in loudnesses is used to calculate an appropriate scaling factor applied to the accompaniment recording such that both recordings have equal loudness. The two recordings are then summed to produce the “full mix” recording used as test input to the proposed system. This technique was used both for the karaoke data and the multi-track data (for which it produced very similar vocal/accompaniment balance to the original mix).

4.2 Results

The system was evaluated in a number of ways. Firstly, informal listening tests were used to determine how successfully the semitone-cancellation system attenuated non-vocal instruments while leaving the lead vocals intact. Next, the two pitch estimation systems were evaluated by considering the accuracy of their estimates during voiced frames. The modified HMM system was then tested with High-Frequency Correlogram (“HF”) estimates only, and then with Semitone-Cancellation/TWM (“SC”) estimates only. Finally the proposed system was tested by running the modified HMM with both sets of pitch estimates.

Figure 4.1 illustrates the output of each of the systems discussed in this chapter for a seven-second region of a test recording. Specifically, the raw pitch estimates from each system, the pitch estimates from each system after HMM post-processing, and the output of the proposed system are shown. The figure illustrates that the raw output from the pitch estimation systems is not itself a suitable transcription, that HMM post-processing can produce a reasonable transcription (depending on the pitch estimates input), and that the proposed system can combine good estimates from both systems to produce a better transcription than either system alone. These points are discussed in detail in what follows.

4.2.1 Informal listening tests for semitone-cancellation

Listening to the output of the semitone-cancellation system for the test recordings verified that while the voice is affected by the procedure, the pitch and lyrics remain mostly intact. The success of the system in attenuating non-vocal notes depends on the instrumentation of the accompaniment. In general bass, guitar and keyboard notes are heavily attenuated (except for their noise-like onsets). The attenuation of bowed strings (which are present in several of the karaoke test recordings) is much lower than other instruments, and some wind instruments are also resilient to the procedure.

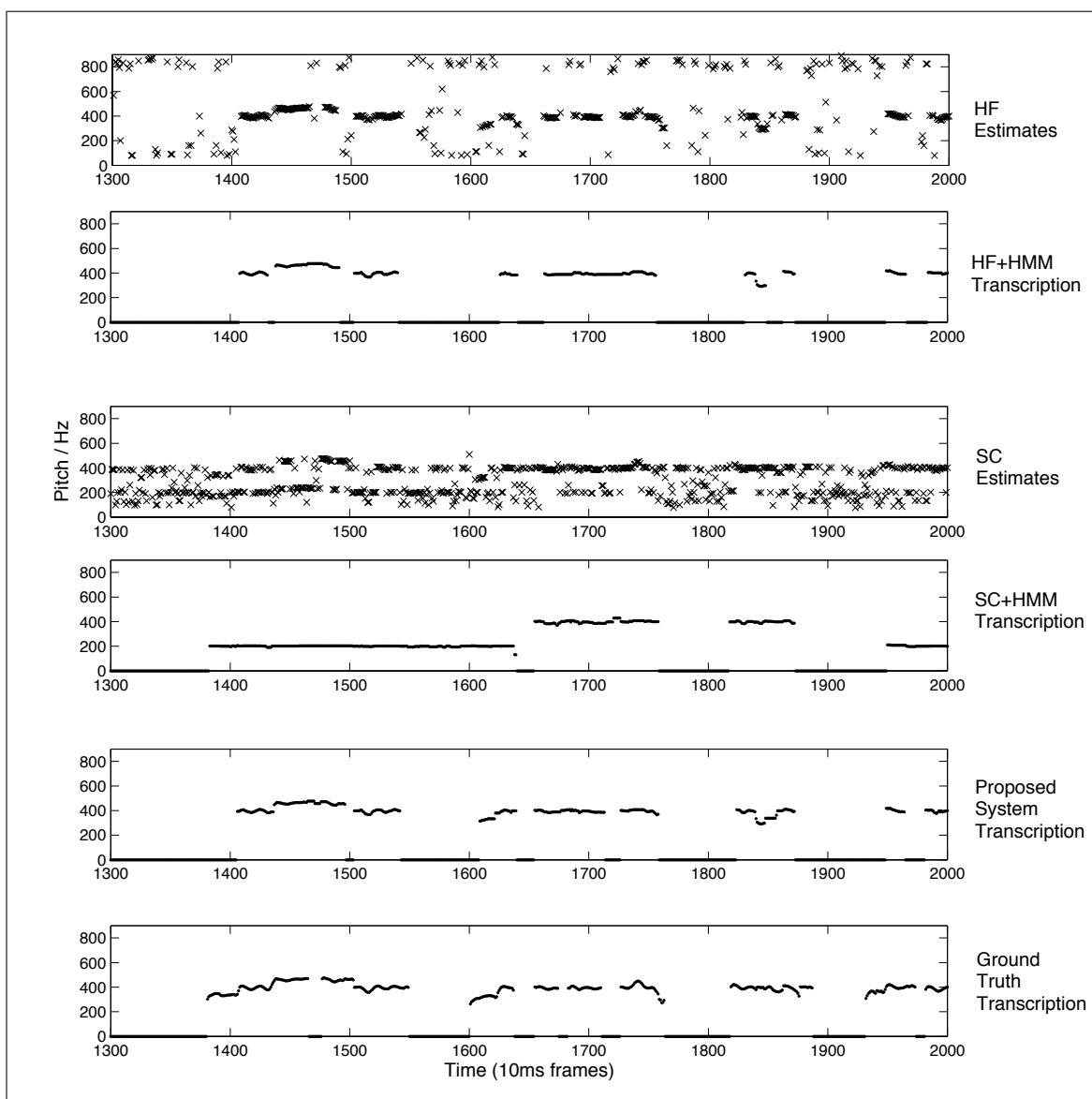


Figure 4.1: Example of the outputs of each pitch estimation system with and without HMM post-processing, the output of the proposed transcription system, and the ground truth melody transcription.

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Mean
HF (%)	75	63	61	46	49	52	50	63	63	44	69	41	43	66	36	31	71	39	40	53
SC (%)	67	48	80	54	41	56	57	26	63	51	65	57	55	67	54	47	68	55	43	55
Either (%)	88	76	88	66	62	70	74	69	81	67	84	68	67	81	66	58	86	68	64	73

Table 4.1: Percentage of estimates correct from the HF system, SC system, and either system.

4.2.2 Accuracy of pitch estimates

The estimates from each pitch estimation system were compared with the ground truth for voiced frames, and the percentage of estimates within 50 cents of the ground truth pitch was recorded. The results are shown in Table 4.1. Also shown is the percentage of voiced frames for which *either* system produced a correct estimate, included to shown the potential gain in accuracy from combining the sets of estimates well.

There are several conclusions to be drawn from these results. It is clear that neither set of raw estimates achieves good transcription accuracy reliably across all (or even most) test recordings. Both systems display a wide range of accuracy scores, achieving high accuracies (around 70–80%) on certain tracks and particularly low accuracies (below 40%) on others. It is important to note the disparity between scores for the two systems — generally one system’s accuracy is at least 10% higher than the other’s, and which is higher depends on the recording. This indicates that higher overall accuracy may be achieved by choosing well between the estimates from the two pitch estimation systems in the HMM system.

The accuracy scores for either system being correct demonstrate another important point. Rather than simply being the higher of the two system scores, the combined score is typically at least 10% higher than either system alone achieves. This shows that the errors made by the two systems are different and increases the potential accuracy gain from correctly selecting estimates in the HMM system.

4.2.3 Accuracy of pitch estimates with HMM post-processing

As was discussed in Section 2.3, the accuracy of pitch estimation systems can often be improved by suitable post-processing on the raw frame estimates, such as by using a HMM to model notes. To measure the potential accuracy of the two pitch estimation systems with suitable post-processing, the HMM designed for the proposed system was applied to a single set of estimates, and the resulting pitch accuracy scores are shown in Table 4.2.

Applying post-processing increases the accuracy score in 27 of the 38 tests, in some cases by 10% or more. In the cases where post-processing reduces accuracy, the reduction is caused by the pitch estimates being so scattered during voiced periods that the HMM fails to track the true pitch.

Test :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Mean
HF (%)	75	63	61	46	49	52	50	63	63	44	69	41	43	66	36	31	71	39	40	53
HF+HMM (%)	83	71	74	54	48	64	58	74	70	56	77	52	46	71	33	36	75	50	31	59
SC (%)	67	48	80	54	41	56	57	26	63	51	65	57	55	67	54	47	68	55	43	55
SC+HMM (%)	75	56	81	58	39	53	63	17	81	37	74	59	46	75	66	33	71	51	23	56

Table 4.2: Percentage of estimates correct from HF and SC systems, with and without HMM post-processing.

These single-estimate HMM systems may also be evaluated according to the criteria of the 2005 MIREX Melody Extraction task, introduced previously in Section 2.4.1 and summarised in the next section. The results of this evaluation are shown in Table 4.4 in the next section, along with the results for the proposed system.

4.2.4 Accuracy of proposed system

The system proposed in this report uses the estimates from both pitch estimation systems as input to the modified HMM to produce a transcription which is hopefully more accurate than either pitch estimation system alone produces. The system is evaluated according to the criteria used in the MIREX 2005 Melody Extraction competition. The system is scored on voicing detection (proportion of voiced frames marked voiced), false alarm rate (proportion of unvoiced frames marked voiced), the d-prime measure (which summarises voicing accuracy), the raw pitch accuracy (proportion of estimates within 50 cents of the ground truth (ignoring voicing)), the raw chroma accuracy (raw pitch accuracy allowing octave errors), and finally its overall accuracy (proportion of frames which are correctly voiced and, in the voiced case, estimated correct to within 50 cents).

The transcription format for the MIREX competition was to transcribe negative pitch estimates to indicate unvoiced frames, so that pitch accuracy could be judged independently of a system’s voicing detection. To produce such a transcription, the proposed system was run twice for each test, once with silent states allowed in the HMM and once without. For frames which the first run marked as unvoiced, estimates from the second run were inserted as negative values. The full results for the proposed system are shown in Table 4.3.

Generally the results are promising. The pitch accuracy is very high for some recordings and the average value of 71% correct estimates is high by the standards of modern melody transcription systems. The voicing scores are also quite good considering the very varied nature of the test set. While a higher d-prime measure was hoped for, it is a strong result for a novel (and relatively simple) V/UV segmentation method.

It can be seen that the pitch accuracy varies quite considerably across performances (though not as much as the individual systems’ pitch accuracy). Three recordings stand out as having poor pitch accuracy scores (numbers 5, 16 and

Test	Voicing Det.	False Alarm Rate	d'	Raw Pitch Acc.	Raw Chroma Acc.	Overall Acc.
1	0.91	0.49	-	0.88	0.93	0.66
2	0.82	0.28	-	0.75	0.81	0.70
3	0.86	0.34	-	0.89	1.00	0.76
4	0.62	0.24	-	0.63	0.67	0.64
5	0.57	0.16	-	0.53	0.59	0.66
6	0.62	0.012	-	0.71	0.74	0.72
7	0.70	0.43	-	0.68	0.77	0.59
8	0.88	0.24	-	0.64	0.78	0.64
9	0.80	0.27	-	0.87	0.90	0.73
10	0.66	0.049	-	0.67	0.76	0.74
11	0.85	0.37	-	0.87	0.89	0.74
12	0.64	0.046	-	0.66	0.74	0.69
13	0.64	0.32	-	0.64	0.66	0.56
14	0.83	0.18	-	0.79	0.82	0.76
15	0.54	0.24	-	0.72	0.85	0.63
16	0.58	0.26	-	0.54	0.58	0.53
17	0.87	0.23	-	0.83	0.87	0.79
18	0.69	0.45	-	0.69	0.72	0.58
19	0.38	0.002	-	0.44	0.53	0.62
All :	0.71	0.24	1.25	0.71	0.77	0.67

Table 4.3: Evaluation of proposed system for 19 30-second test recordings using MIREX evaluation criteria.

19). In the case of recording 5 (“Hello, Dolly”) the vocalist sings in the style of Louis Armstrong, and the low pitch accuracy score is likely caused by the distinctive singing style. A predominant F0 transcription system (see next section) also exhibited a particularly low score (34%) on this recording. The low score on recording 16 (eighties pop song “Machinery”) is quite surprising, as the vocals seem quite clear to a human listener. It is hard to know which of the several accompaniment instruments is problematic. Again, the predominant F0 method also performed quite poorly (43%) on this track. Finally, track 19 (“The Hand That Feeds”) is one of two Nine Inch Nails tracks used to test the system, and it is likely that the persistent percussion causes difficulty for the pitch estimators. Interestingly the predominant F0 method seems more resilient to the percussive noise, achieving 65% pitch accuracy on this track. These three tracks aside, the proposed system displays strong pitch accuracy scores (60–90%).

The scores for chroma accuracy are roughly 5–10% higher than those for pitch, indicating the system does suffer from octave errors. This was somewhat surprising since octave errors hadn’t been common in the system’s performance on the training set. However, the average pitch accuracy is similar to that for the training set which suggests that the system is flexible with regards to its input data and hasn’t been overtrained.

The voicing scores are somewhat lower than hoped. This could be partly due to the test set exhibiting longer and more frequent unvoiced periods than the training set, but since both voicing detection and the false alarm rate display a wide range of values, it is unlikely to be a simple V/UV balance issue. The low voicing detection

rates are most likely due to regions in recordings in which neither pitch estimation system reliably produces correct estimates. This could be tackled by trying to improve one or both of the pitch estimation systems, or by adding a third pitch estimation system — perhaps a predominant F0 system with low associated reliability scores. Naturally there is a risk that this would lead to higher false alarm rates also, but if additional pitch estimation systems are well designed and complement the current two systems, more accurate V/UV segmentation should be possible.

The high false alarm rates are concerning as they suggest that the system is transcribing notes from other instruments during vocal silence. Preliminary investigation suggests that the two pitch estimation systems are responsible for approximately equal numbers of false alarm estimates. That is, about half the superfluous voiced frames are supported by HF pitch estimates and about half by SC pitch estimates.

It was hoped that the scattering of estimates during UV regions would permit accurate V/UV segmentation without further considerations. However, there are possible additions which could improve the segmentation.

Currently the pitch estimation systems take no account of the absolute energy of the pitch candidates they select. For example, the TWM algorithm will favour pitches whose measured peaks near the predicted frequencies are high, but in the absence of a loud harmonic sound, will choose a weak harmonic sound. This means that notes which survive the semitone-cancellation procedure could very well be transcribed when the vocals are silent. A simple energy threshold (perhaps based on the measured energies of previous estimates for the current recording) could be used to determine whether a pitch estimate is given, or silence is transcribed. This would work in conjunction with the current scattering of estimates during UV regions. The system by Poliner and Ellis[23] showed that a simple energy thresholding system can achieve good segmentation accuracy, so such an addition to the proposed system could potentially greatly improve V/UV segmentation.

The overall scores are generally lower than the raw pitch accuracy scores, indicating that the voicing detection is currently the weaker part of the system. However, the average overall score is not very far removed from the average pitch accuracy score, so the voicing performance is not substantially worse.

The proposed system may also be compared with the systems from the previous section (which use a single pitch estimation method with HMM post-processing) to indicate whether using multiple pitch estimation systems is advantageous. The summary measures for each system are shown in Table 4.4.

The proposed system displays considerably higher raw pitch accuracy than either single-estimate system. This is evidence that the HMM reliably chooses the correct estimate when one is available. This can be credited to the reliability measures associated with estimates, and perhaps also to the transition probability model (since incorrect estimates from a system may be scattered in frequency and the associated states will have low transition probabilities).

System	Voicing Det.	False Alarm Rate	d'	Raw Pitch Acc.	Raw Chroma Acc.	Overall Acc.
HF+HMM	0.58	0.17	1.20	0.59	0.63	0.63
SC+HMM	0.68	0.29	1.00	0.56	0.67	0.58
Proposed	0.71	0.24	1.25	0.71	0.77	0.67
Modified YIN	-	-	-	0.56	0.67	-

Table 4.4: Summary evaluation using MIREX evaluation criteria for 19 30-second test recordings. Scores are shown for the two single-estimate HMM systems, the proposed system, and a predominant F0 transcription system.

The proposed system also exhibits better V/UV segmentation than either single-estimate system (though not significantly in the HF case). This is promising since poor voicing from one system is not propagating into the combined system. This supports the idea that adding further pitch estimation systems could improve V/UV segmentation (as suggested above).

4.2.5 Comparison with existing systems

Since the data set used in the tests above is different from that used in the 2005 MIREX Melody Extraction competition, the proposed system’s performance cannot be directly compared with the scores obtained by systems in that competition. However, since the test data set used here is reasonably large and varied, the scores should be indicative of the system’s likely performance on the MIREX data, and a rough comparison can be made.

The systems which performed V/UV segmentation in the competition achieved d-prime measures ranging from 0.83 to 1.85. This suggests that although the proposed system’s voicing performance is not necessarily an improvement over previous methods, it is reasonably competitive.

The raw pitch accuracy scores in the competition ranged from about 60% to 70%. Again, the difference in data sets prevents any direct comparison, but the proposed system could reasonably be expected to exhibit pitch accuracy in that same range on the MIREX data set.

Assuming the voicing and pitch estimation performance of the proposed system were similar on the MIREX data set as on the data set used here for evaluation, the overall score would be among those of the top scoring systems in the competition (overall scores ranged from around 60% to 70%).

For the sake of having some point of direct comparison, one of the predominant F0 systems entered in the competition (described in Section 3.2 of [30]) was used to transcribe the test data. It is a modified version of the YIN algorithm with perceptual weighting of frequency candidates. The algorithm achieved raw pitch accuracy of around 60% in the MIREX competition. The summary measures for its performance on the test data are shown in Table 4.4.

The modified YIN approach exhibits similar pitch accuracy to the two single-estimate systems and substantially lower accuracy than the proposed system. It is difficult to say whether the higher accuracy of the improved system stems solely from using multiple pitch-estimation systems together or if it is important that the pitch estimation systems are designed specifically for vocal melodies. That the YIN approach’s performance on the test set is similar to its performance in the MIREX competition supports the suggestion that the proposed system could perform well on the MIREX data set.

4.3 Future Work

There are several possible extensions to this work. Although V/UV segmentation performance is reasonable, the restriction to vocal melodies should permit better accuracy than generic melody transcription systems achieve and so there is room for improvement. The chosen approach of relying on pitch estimates becoming scattered during vocal silence seems to work quite well, but adding a threshold on the estimates such that the two pitch estimation systems do transcribe silence could greatly improve the overall V/UV segmentation.

For the SC/TWM system this could be implemented by consideration of the spectral energy of the harmonics of the chosen pitch, as described in Section 4.2.4 above. For the HF/Correlogram system, a threshold could be set on the number of channels supporting an estimate, such that if the cluster with highest population only has a population of one or two channels, silence is transcribed. Alternatively some consideration of the energy in the high-frequency channels could be used to judge whether a voice is present.

The pitch accuracy achieved is promising, but again, it was hoped that restricting the task to vocal melodies would permit higher accuracies than those achieved by general systems. The prevalence of octave errors for the test set was a surprising result, and could perhaps be addressed by special consideration of octave relationships in the observation and transition probabilities of the HMM.

Although two characteristics of the human voice are used to tailor the pitch estimation systems to vocal melodies, more established timbre cues (such as those proposed for the MPEG-7 standard[47] or those previously used for instrument recognition in a polyphonic context[48]) are not used by the proposed system. There are two ways such cues could be used in the proposed system. One approach would be to learn suitable timbre feature models for the human singing voice, and exclude pitch estimates/regions which are unlikely under that model. Alternatively, once features are computed for the current recording the distributions of observed feature values could be analysed, and on the assumption that most transcribed notes are vocal, outliers in feature space could be excluded to remove non-vocal notes. Using timbre features in either of these ways could potentially improve both pitch and voicing accuracy.

Two pitch estimation systems were used in the proposed system. The results suggest that the approach of using multiple pitch estimation systems can greatly

improve upon the accuracies of the single systems. If additional pitch estimation systems could be designed which complement the existing two (ie. have different error sets), it would be interesting to see whether the overall system's accuracy improves when using more than two pitch estimators.

As discussed in Section 2.4.1, some transcription systems output a "piano roll" type score rather than exact pitch values. This is a higher level representation more akin to a MIDI recording or real musical score. It would be quite straightforward to apply a system such as Rynänen and Klapuri's note-modelling HMM for singing transcription[18] to the output of the proposed system to produce such a score.

While the sections above offer some context for judging the performance of the proposed system, there can be no substitute for a direct comparison with other methods. The proposed system will therefore be entered in this year's MIREX Melody Extraction competition, to be tested only on those recordings whose melody is sung. The test data set to be used in the 2006 competition is the same as that used in 2005, so the results will be comparable both with the systems entered this year, and those from last year which were discussed in the Background chapter to this report.

5. CONCLUSIONS

As set out in the introduction to this report, the main aim of this project was to design and build a system to transcribe the vocal melody in popular music recordings. The MIREX[10] Melody Extraction competition in 2005 demonstrated the difficulty faced by melody transcription systems of correctly distinguishing melody from accompaniment. The system proposed here is designed to avoid such ambiguities by seeking to transcribe sung notes only, and hence produce a transcription of the vocal melody. It was thought that this restriction would allow higher transcription accuracy than was demonstrated by the general melody transcription systems in the MIREX competition.

One existing system for vocal melody transcription in a polyphonic context was found[11], and the basis for its vocal specificity was investigated as part of this project. Experiments confirmed that the human voice dominates over other instruments in high frequencies (3–15kHz). This led to the development of a high-frequency correlogram as one of two pitch estimation systems used by the proposed transcription system. The second pitch estimation system is based on a semitone-cancellation procedure which was adapted from a technique previously used for vocal detection[13].

The estimates from these two pitch estimation systems were combined by a modified Hidden Markov Model to produce the final vocal melody transcription. The proposed model differs from standard approaches in several ways, most notably in taking input from multiple distinct pitch estimation methods, transcribing a continuous pitch contour, and inferring silence from scattered estimates.

A set of recordings and transcriptions was assembled to test the system with a variety of musical styles and instrumentations. The results are very promising, with pitch accuracy comparable to the top scores of systems in the 2005 MIREX Melody Extraction competition. Voicing performance is good (especially since the approach is novel) and there are several possibilities discussed in the previous chapter for extending this approach and improving accuracy.

Further work and testing would be necessary to determine whether restricting the melody transcription task to vocal melodies truly permits higher accuracy transcription. However, the high accuracy scores achieved by the proposed system certainly encourage further investigation of vocal melody transcription.

Additionally, the results show that combining estimates from multiple pitch estimation systems can produce considerably higher accuracy than any of the individual systems (if the pitch estimation systems are well chosen to complement each other). The modified HMM proposed here is shown to be a good way of combining such estimates and the technique may be effective for other types of transcription system.

The proposed system will be entered in this year's MIREX Melody Extraction competition to be tested alongside state of the art melody transcription systems.

References

- [1] Musical Instrument Digital Interface. Information available from the MIDI Manufacturers Association, <http://www.midi.org/>.
- [2] Samer A. Abdallah and Mark D. Plumbley. Unsupervised analysis of polyphonic music by sparse coding. *IEEE Transactions on Neural Networks*, 17(1), January 2006.
- [3] Matti P. Rynnänen and Anssi Klapuri. Polyphonic music transcription using note event modeling. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.
- [4] M. Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6:439–449, June 2004.
- [5] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111, 2002.
- [6] Robert C. Maher and James W. Beauchamp. Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *Journal of the Acoustical Society of America*, 95(4):2254–2263, 1994.
- [7] Malcolm Slaney and Richard F. Lyon. A perceptual pitch detector. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 357–360, 1990.
- [8] D. FitzGerald, B. Lawlor, and E. Coyle. Drum transcription in the presence of pitched instruments using prior subspace analysis. In *Proceedings of the Irish Signals and Systems Conference*, July 2003.
- [9] Description of MIREX 2005 Melody Extraction task. http://www.music-ir.org/mirex2005/index.php/Audio_Melody_Extraction.
- [10] Stephen J. Downie, Kris West, Andreas Ehmann, and Emmanuel Vincent. The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005) : Preliminary Overview. In *Proceedings of the sixth international conference on music information retrieval (ISMIR)*, 2005.
- [11] Yipeng Li and DeLiang Wang. Detecting pitch of singing voice in polyphonic audio. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.
- [12] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77 issue 2, pages 257–286, 1989.
- [13] Arun Shenoy, Yuansheng Wu, and Ye Wang. Singing voice detection for karaoke application. In *Visual Communications and Image Processing (VCIP)*, 2005.

- [14] Udo Zölzer, editor. *DAFX : Digital Audio Effects*. John Wiley & Sons, 2002. ISBN: 0-471-49078-4.
- [15] Matthew E. P. Davies and Mark D. Plumbley. Causal tempo tracking of audio. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*, volume 3, pages 164–169, 2004.
- [16] R. O. Duda, R. F. Lyon, and M. Slaney. Correlograms and the separation of sounds. In *Proceedings of the Twenty-Fourth Asilomar Conference on Signals, Systems and Computers*, volume 1, page 457, 1990.
- [17] S. J. Young, N. H. Russell, and J. H. S. Thornton. Token passing: A simple conceptual model for connected speech recognition systems. Technical report, Cambridge University, 1989. F-INFENG/TR38.
- [18] Matti P. Rynnänen and Anssi Klapuri. Modelling of note events for singing transcription. In *Workshop on Statistical and Perceptual Audio Processing (SAPA)*, 2004.
- [19] Christopher Harte and Mark Sandler. Automatic Chord Identification Using a Quantised Chromagram. In *Proceedings of the Audio Engineering Society 118th Convention*, 2005.
- [20] Masataka Goto. A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages II-757–760, June 2000.
- [21] Karin Dressler. Extraction of the melody pitch contour from polyphonic audio. In *Proceedings of the 2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2005. Available at <http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html>.
- [22] Matti Rynnänen and Anssi Klapuri. Note event modelling for audio melody extraction. In *Proceedings of the 2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2005. Available at <http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html>.
- [23] Graham E. Poliner and Daniel P. W. Ellis. A classification approach to melody transcription. In *Proceedings of the 2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2005. Available at <http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html>.
- [24] Training data for MIREX 2005 Melody Extraction competition. Currently available at <http://labrosa.ee.columbia.edu/projects/melody/>.
- [25] Evaluation of MIREX 2005 Melody Extraction competition. <http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html>.
- [26] Paul Brossier. Fast melody extraction using AUBIO. In *Proceedings of the 2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2005. Available at <http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html>.

- [27] Masataka Goto. PreFEst: A Predominant-F0 Estimation Method for Polyphonic Musical Audio Signals. In *Proceedings of the 2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2005. Available at <http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html>.
- [28] Matija Marolt. Audio melody extraction based on timbral similarity. In *Proceedings of the 2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2005. Available at <http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html>.
- [29] Rui Pedro Paiva. An algorithm for melody detection in polyphonic recordings. In *Proceedings of the 2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2005. Available at <http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html>.
- [30] Emmanuel Vincent and Mark D. Plumbley. Predominant-F0 estimation using Bayesian harmonic waveform models. In *Proceedings of the 2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2005. Available at <http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html>.
- [31] John Daugman. Biometric decision landscapes. Technical Report 482, University of Cambridge Computer Laboratory, January 2000.
- [32] Ixone Arroabarren and Alfonso Carlosena. Voice production mechanisms of vocal vibrato in male singers. *IEEE Transactions on Audio, Speech and Language Processing*, 2006. Accepted for future publication.
- [33] R. Timmers and P. W. M. Desain. Vibrato: The questions and answers from musicians and science. In *Proceedings of the sixth International Conference on Music Perception and Cognition*, 2000.
- [34] Malcolm Slaney. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. Technical Report 35, Apple Computer Inc., 1993. Available from.
- [35] M. Wu, D. L. Wang, and G. J. Brown. A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, 11:229–241, 2003.
- [36] X. Amatriain, J. Bonada, A. Loscos, and X. Serra. *DAFX : Digital Audio Effects*, chapter 10 : Spectral Processing. John Wiley & Sons, 2002. ISBN: 0-471-49078-4.
- [37] M. Lagrange, S. Marchand, and J.-B. Rault. Tracking partials for the sinusoidal modeling of polyphonic sounds. In *The IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, March 2005.
- [38] Malcolm Slaney. Auditory toolbox. Technical Report 010, Interval Research Corporation, 1998. Available from <http://www.slaney.org/malcolm/pubs.html>.

- [39] N. Ratnayake, M. Savic, and J. Sorensen. Use of semi-Markov models for speaker-independent phoneme recognition. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 565–568, March 1992.
- [40] Paul Boersma and David Weenink. Praat: doing phonetics by computer (Version 4.4.25). Available from <http://www.praat.org/>.
- [41] M. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley and Sons, 1996. ISBN: 0471594318.
- [42] BASS-dB: the Blind Audio Source Separation evaluation database. <http://bass-db.gforge.inria.fr/BASS-dB/>.
- [43] Creative Commons licenses. <http://creativecommons.org/>.
- [44] Peter Gabriel remix contest on Real World Remixed, September 2006. <http://www.realworldremixed.com/>.
- [45] Nine Inch Nails multi-track downloads. <http://www.nin.com/access/only/>.
- [46] Apple Computer Inc. Garageband version 3. <http://www.apple.com/ilife/garageband/>.
- [47] G. Peeters, S. McAdams, and P. Herrera. Instrument Sound Description in the Context of MPEG-7. In *Proceedings of the International Computer Music Conference (ICMC)*, September 2000.
- [48] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 245–248, March 2005.