EURASIP Journal on Advances in Signal Processing

SpringerOpen

Efficient methods for joint estimation of multiple fundamental frequencies in music signals

EURASIP Journal on Advances in Signal Processing 2012,
2012:27 doi:10.1186/1687-6180-2012-27

Antonio Pertusa (pertusa@dlsi.ua.es)
Jose M. Inesta (inesta@dlsi.ua.es)

ISSN 1687-6180

Article type Research

Submission date 11 April 2011

Acceptance date 14 February 2012

Publication date 14 February 2012

Article URL http://asp.eurasipjournals.com/content/2012/1/27

Efficient methods for joint estimation of multiple fundamental frequencies in music signals

Antonio Pertusa* and José M Iñesta

Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante,

P.O. Box 99, E-03080 Alicante, Spain

*Corresponding author: pertusa@dlsi.ua.es

Email address:

JMI: inesta@dlsi.ua.es

Abstract

This study presents efficient techniques for multiple fundamental frequency estimation in music signals. The proposed methodology can infer harmonic patterns from a mixture considering interactions with other sources and evaluate them in a joint estimation scheme. For this purpose, a set of fundamental frequency candidates are first selected at each frame, and several hypothetical combinations of them are generated. Combinations are independently evaluated, and the most likely is selected taking into account the intensity and spectral

smoothness of its inferred patterns. The method is extended considering adjacent frames in order to smooth the detection in time, and a pitch tracking stage is finally performed to increase the temporal coherence. The proposed algorithms were evaluated in MIREX contests yielding state of the art results with a very low computational burden.

---

1 Introduction

The goal of a multiple fundamental frequency (f_0) estimation method is to infer the number of simultaneous harmonic sounds present in an acoustic signal and their fundamental frequencies. This problem is relevant in speech processing, structural audio coding, and several music information retrieval (MIR) applications, like automatic music transcription, compression, instrument separation and chord estimation, among others.

In this study, a multiple f_0 estimation method is presented for the analysis of pitched musical signals. The core methodology introduced in [1] is described and extended considering information about neighbor frames.

Most multiple f_0 estimation methods are complex systems. The decomposition of a signal into multiple simultaneous sounds is a challenging task due to harmonic overlaps and inharmonicity (when partial frequencies are not exact multiples of the f_0). Many different techniques are proposed in the literature to face this task. Recent reviews of multiple f_0 estimation in music signals can be found in [2–4].

Some techniques rely on the mid-level representation, trying to emphasize the

underlying fundamental frequencies by applying signal processing transformations to the input signal [5–7]. Supervised [8, 9] and unsupervised [10, 11] learning techniques have also been investigated for this task. The matching pursuit algorithm, which approximates a solution for decomposing a signal into linear functions (atoms), is also adopted in some approaches [12, 13]. Methods based on statistical inference within parametric signal models [3, 14, 15] have also been studied for this task.

Heuristic approaches can also be found in the literature. Iterative cancellation methods estimate the prominent f_0 subtracting it from the mixture and repeating the process until a termination criterion [16–18]. Joint estimation methods [19–21] can evaluate a set of possible f_0 hypotheses, consisting of f_0 combinations, selecting the most likely at each frame without corrupting the residual as it occurs with iterative cancellation.

Some existing methods can be switched to another framework. For example, iterative methods can be viewed against matching pursuit background, and many unsupervised learning methods like [11] can be switched to a statistical framework.

Statistical inference provides an elegant framework to deal with this problem, but these methods are usually intended for single instrument f_0 estimation (typically piano), as exact inference often becomes computationally intractable for complex and very different sources.

Similarly, supervised learning methods can infer models of pitch combinations seen in the training stage, but they are currently constrained to monotimbral sounds with almost constant spectral profiles [4].

In music, consonant chords include harmonic components of different sounds which coincide in some of their partial frequencies (harmonic overlaps). This situation is very frequent and introduces ambiguity in the analysis, being the main challenge in multiple f_0 estimation. When two harmonics are

overlapped, two sinusoids of the same frequency are summed in the waveform, resulting a signal with the same frequency and which magnitude depends on their phase difference.

The contribution of each harmonic to the mixture can not be properly estimated without considering the interactions with the other sources. Joint estimation methods provide an adequate framework to deal with this problem, as they do not assume that sources are mutually independent and individual pitch models can be inferred taking into account their interactions. However, they tend to have high computational costs due to the number of possible combinations to be evaluated.

Novel efficient joint estimation techniques are presented in this study. In contrast to previous joint approaches, the proposed algorithms have a very low computational cost. They were evaluated and compared to other studies in MIREX [22, 23] multiple f_0 estimation and tracking contests, yielding competitive results with very efficient runtimes.

The core process, introduced in [1], relies on the inference and evaluation of spectral patterns from the mixture. For a proper inference, source interactions must be considered in order to estimate the amplitudes of their overlapped harmonics. This is accomplished by evaluating independent combinations consisting of hypothetical patterns (f_0 candidates). The evaluation criterion enhances those patterns having high intensity and smoothness. This way, the method takes advantage of the spectral properties of most harmonic sounds, in which first harmonics are usually those with higher energy and their spectral profile tend to be smooth.

Evaluating many possible combinations can computationally intractable. In this study, the efficiency is boosted by reducing the spectral information to be considered for the analysis, adding a f_0 candidate selection process, and pruning unlikely combinations by applying some constraints, like a minimum

intensity for a pattern.

One of the main contributions of this study is the extension of the core algorithm to increase the temporal coherence. Instead considering isolated frames, the combinations sharing the same pitches across neighbor frames are grouped to smooth the detection in time. A novel pitch tracking stage is finally presented to favor smooth transitions of pitch intensities.

The proposed algorithms are publicly available at http://grfia.dlsi.ua.es/cm/projects/drims/software.php.

The overall scheme of the system can be seen in Figure 1. The core methodology performing a frame by frame analysis is described in Sec. 2, whereas the extended method which considers temporal information is presented in Sec. 3. The evaluation results are described in Sec. 4, and the conclusions and perspectives are finally discussed in Sec. 5.

2 Methodology

Joint estimation methods generate and evaluate competing sets of f_0 combinations in order to select the most plausible combination directly. This scheme, recently introduced in [24, 25] has the advantage that the amplitudes of overlapping partials can be approximated taking into account the partials of the other candidates for a given combination. Therefore, partial amplitudes can depend on the particular combination to be evaluated, opposite to an iterative estimation scheme like matching pursuit, where a wrong estimate may produce cumulative errors.

The core method performs a frame by frame analysis, selecting the most likely combination of fundamental frequencies at each instant. For this purpose, a set of f_0 candidates are first identified from the spectral peaks. Then, a set of possible combinations, $\mathcal{C}(t)$, of candidates are generated, and a joint algorithm

is used to find the most likely combination.

In order to evaluate a combination, hypothetical partial sequences HPS (term proposed in [26] to refer to a vector containing hypothetical partial amplitudes) are inferred for its candidates. In order to build these patterns, harmonic interactions with the partials of the other candidates in the combination are considered. The overlapped partials are first identified, and their amplitudes are estimated by linear interpolation using the non-overlapped harmonic amplitudes.

Once patterns are inferred, they are evaluated taking into account the sum of its hypothetical harmonic amplitudes and a novel smoothness measure. Combinations are analysed considering their individual candidate scores, and the most likely combination is selected at the target frame.

The method assumes that the spectral envelopes of the analysed sounds tend to vary smoothly as a function of frequency. The spectral smoothness principle has successfully been used in different ways in the literature [7, 26–29]. A novel smoothness measure based on the convolution of the hypothetical harmonic pattern with a Gaussian window is proposed.

The processing stages, shown in Figure 1, are described below.

2.1 Preprocessing

The analysis is performed in the frequency domain, computing the magnitude spectrogram using a 93 ms Hanning windowed frame with a 9.28 ms hop size. This is the frame size typically chosen for multiple f_0 estimation of music signals in order to achieve a suitable frequency resolution, and it experimentally showed to be adequate. The selected frame overlap ratio may seem high from a practical point of view, but it was required to compare the method with other studies in MIREX (see 4.3).

To get a more precise estimation of the lower frequencies, zero padding is used multiplying the original window size by a factor z to complete it with zeroes before computing the FFT.

In order to increase the efficiency, many unnecessary spectral bins are discarded for the subsequent analysis using a simple peak picking algorithm to extract the hypothetical partials. At each frame, only those spectral peaks with an amplitude higher than a threshold μ are selected, removing the rest of spectral information and obtaining this way a sparse representation containing a subset of spectral bins. It is important to note that this thresholding does not have a significant effect on the results, as values of μ are quite low, but the efficiency of the method importantly increases.

2.2 Candidate selection

The evaluation of all possible f_0 combinations in a mixture is computationally intractable, therefore a reduced subset of candidates must be chosen before generating their combinations. For this, candidates are first selected from the spectral peaks within the range $[f_{\min}, f_{\max}]$ corresponding to the musical pitches of interest. Harmonic sounds with missing fundamentals are not considered, although they seldom appear in practical situations. A minimum spectral peak amplitude ε for the first partial (f_0) can also be assumed in this stage.

The spectral magnitudes at the candidate partial positions are considered as a criterion for candidate selection as described next.

2.2.1 Partial search

Slight harmonic deviations from ideal partial frequencies are common in music sounds, therefore inharmonicity must be considered for partial search. For

this, a constant margin around each harmonic frequency $f_h \pm f_r$ is set. If there are no spectral peaks within this margin, the harmonic is considered to be missing. Besides considering a constant margin, frequency dependent margins were also tested assuming that partial deviations in high frequencies are larger than those in low frequencies. However, results decreased, mainly because many false positive harmonics (most of them corresponding to noise) can be found in high frequencies.

Different strategies were also tested for partial search, and finally, like in [30], the harmonic spectral location and spectral interval principles [31] were chosen in order to take inharmonicity into account. The ideal frequency f_h of the first harmonic is initialized to $f_h = 2f_0$. The next ones are searched at $f_{h+1} = (f_x + f_0) \pm f_r$, where $f_x = f_i$ if the previous harmonic h was found at the frequency f_i, or $f_x = f_h$ if the previous partial was missing.

In many studies, the closest peak to f_h within a given region is identified as a partial. A novel variation which experimentally slightly increased (although not significantly) the proposed method performance is the inclusion of a triangular window. This window, centered in f_h with a bandwidth $2f_r$ and a unity amplitude, is used to weight the partial magnitudes within this range (see Figure 2). The spectral peak with maximum weighted value is selected as a partial. The advantage of this scheme is that low amplitude peaks are penalized and, besides the harmonic spectral location, intensity is also considered to correlate the most important spectral peaks with partials.

2.2.2 Selection of F candidates

Once the hypothetical partials for all possible candidates are searched, candidates are ordered decreasingly by the sum of their amplitudes and, at most, only the first F candidates of this ordered list are chosen for the

following processing stages.

Harmonic summation is a simple criterion for candidate selection, and other alternatives can be found in the literature, including harmonicity criterion [30], partial beating [30], or the product of harmonic amplitudes in the power spectrum [20]. Evaluating alternative criteria for candidate selection is left as future study.

2.3 Generation of candidate combinations

All the possible combinations of the F selected candidates are calculated and evaluated, and the combination with highest score is yielded at the target frame. The combinations consist of different number of fundamental frequencies. In contrast to studies like [26], there is not need for *a priori* estimation of the number of concurrent sounds before detecting the fundamental frequencies, and the polyphony is implicitly calculated in the f_0 estimation stage, choosing the combination with highest score independently from the number of candidates.

At each frame t, a set of combinations $\mathcal{C}(t) = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$ is obtained. For efficiency, like in [20], only the combinations with a maximum polyphony P are generated from the F candidates. The amount of combinations without repetition (N) can be calculated as:

$$N = \sum_{n=1}^{P} \binom{F}{n} = \sum_{n=1}^{P} \frac{F!}{n!(F-n)!} \tag{1}$$

Therefore, N combinations are evaluated at each frame, so the adequate selection of F and P is critical for the computational efficiency of the algorithm. An experimental discussion on this issue is presented in Sec. 4.2.

2.4 Evaluation of combinations

In order to evaluate a combination $\mathcal{C}_i \in \mathcal{C}(t)$, a hypothetical pattern is first estimated for each of its candidates. Then, these patterns are evaluated in terms of their intensity and smoothness, assuming that music sounds have a perceivable intensity and their spectral shapes are smooth, like it occurs for most harmonic instruments. The combination $\hat{\mathcal{C}}(t)$ which patterns maximize these measures is yielded at the target frame t.

2.4.1 Inference of hypothetical patterns

The intention of this stage is to infer harmonic patterns for the candidates. This is performed taking into account the interactions with other candidates in the analysed combination, assuming that they have smooth spectral envelopes. A pattern (HPS) is a vector $\mathbf{p_c}$ estimated for each candidate $c \in \mathcal{C}$ consisting of the hypothetical harmonic amplitudes of the first H harmonics:

$$\mathbf{p}_c = (p_{c,1}, p_{c,2}, \ldots, p_{c,h}, \ldots, p_{c,H})^{\mathrm{T}} \tag{2}$$

where $p_{c,h}$ is the amplitude for the h harmonic of the candidate c. The partials are searched the same way as previously described for the candidate selection stage. If a particular harmonic is not found within the search margin, then the corresponding value $p_{c,h}$ is set to zero. As in music sounds the first harmonics are usually the most representative and they contain most of the sound energy, only the first H partials are considered to build the patterns. Once the partials of a candidate are identified, the HPS values are estimated taking into account the hypothetical source interactions. For this task, their harmonics are identified and labeled with the candidate they belong to (see Figure 3). After the labeling process, some harmonics will only belong to one candidate (non-overlapped harmonics), whereas others will belong to more

than one candidate (overlapped harmonics).

Assuming that interactions between non-coincident partials (beating) do not alter significantly the original spectral amplitudes, the non-overlapped amplitudes are directly assigned to the HPS. However, the contribution of each source to an overlapped partial amplitude must be estimated.

Getting an accurate estimate of the amplitudes of colliding partials is not reliable only with the spectral magnitude information. In this study, the additivity of linear spectrum is assumed as in most approaches in the literature. Assuming additivity and spectral smoothness, the amplitudes of overlapped partials can be estimated similarly to [26,32] by linear interpolation of the neighboring non-overlapped partials, as shown in Figure 3 (bottom).

If there are two or more consecutive overlapped partials, then the interpolation is done the same way using the available non-overlapped values. For instance, if harmonics 2 and 3 of a pattern are overlapped, then the amplitudes of harmonics 1 and 4 are used to estimate them by linear interpolation.

After the interpolation, the estimated contribution of each partial to the mixture is subtracted before processing the next candidates. This calculation (see Figure 3) is done as follows:

- If the interpolated (expected) value is greater than the corresponding overlapped harmonic amplitude, then $p_{c,h}$ is set as the original harmonic amplitude, and the spectral peak is completely removed from the residual, setting it to zero for the candidates that share that partial.
- If the interpolated value is smaller than the corresponding overlapped harmonic amplitude, then $p_{c,h}$ is set as the interpolated amplitude, and this value is linearly subtracted for the candidates that share the harmonic.

The residual harmonic amplitudes after this process are iteratively analysed

for the rest of the candidates in the combination in ascending frequency order.

2.4.2 Candidate evaluation

The intensity $l(c)$ of a candidate c is a measure of the strength of a source obtained by summing its HPS amplitudes:

$$l(c) = \sum_{h=1}^{H} p_{c,h} \tag{3}$$

Assuming that a pattern should have a minimum loudness, those combinations having any candidate with a very low absolute ($l(c) < \eta$) or relative ($l(c) < \gamma L_{\mathcal{C}}$, being $L_{\mathcal{C}} = \max_{\forall c}\{l(c)\}$) intensity are discarded.

The underlying hypothesis assumes that a smooth spectral pattern is more probable than an irregular one. This is assessed through a novel smoothness measure $s(c)$ which is based on Gaussian smoothing.

To compute it, the HPS of a candidate is first normalized dividing the amplitudes by its maximum value, obtaining $\bar{\mathbf{p}}$. The aim is to compare $\bar{\mathbf{p}}$ with a smooth model $\tilde{\mathbf{p}}$ built from it, in such a way that the similarity between $\bar{\mathbf{p}}$ and $\tilde{\mathbf{p}}$ will give an estimation of the smoothness.

For this purpose, $\bar{\mathbf{p}}$ is smoothed using a truncated normalized Gaussian window $\mathcal{N}_{0,1}$, which is convolved with the HPS to obtain $\tilde{\mathbf{p}}$:

$$\tilde{\mathbf{p}}_c = \mathcal{N}_{0,1} * \bar{\mathbf{p}}_c \tag{4}$$

Only three components were chosen for the Gaussian window of unity variance, $\mathcal{N}_{0,1} = (0.21, 0.58, 0.21)^{\mathrm{T}}$, due to the small size of $\mathbf{p_c}$, which is limited by H. Typical values for H are within the range $H \in [5, 20]$, as only the first harmonics contain most of the energy of a harmonic source.

Then, as shown in Figure 4, a roughness measure $r(c)$ is computed by summing up the absolute differences between $\tilde{\mathbf{p}}$ and the actual normalized

HPS amplitudes:

$$r(c) = \sum_{h=1}^{H} |\tilde{\mathbf{p}}_{c,h} - \bar{\mathbf{p}}_{c,h}| \tag{5}$$

The roughness $r(c)$ is normalized into $\bar{r}(c)$ to make it independent of the intensity:

$$\bar{r}(c) = \frac{r(c)}{1 - \mathcal{N}_{0,1}(\bar{x})} \tag{6}$$

And finally, the smoothness $s(c) \in [0, 1]$ of a HPS is calculated as:

$$s(c) = 1 - \frac{\bar{r}(c)}{H_c} \tag{7}$$

where H_c is the index of the last harmonic found for the candidate. This factor was introduced to prevent that high frequency candidates that have less partials than those at low frequencies will have higher smoothness. This way, the smoothness is considered to be more reliable when there are more partials to estimate it.

A candidate score is computed taking into account the HPS smoothness and intensity:

$$S(c) = l(c) \cdot s^{\kappa}(c) \tag{8}$$

where κ is a factor that permits to balance the smoothness contribution experimentally.

2.4.3 Combination selection

Once all candidates are evaluated, a salience measure $S(\mathcal{C}_i)$ for a combination $\mathcal{C}_i$ is computed as:

$$S(\mathcal{C}_i) = \sum_{c=1}^{|\mathcal{C}|} [S(c)]^2 \tag{9}$$

When there are overlapped partials, their amplitudes are estimated by interpolation, therefore the HPS smoothness tends to increase. To partially compensate this effect in $S(\mathcal{C}_i)$, the candidate scores are squared in order to boost the highest values. This favors a sparse representation, as it is convenient to explain the mixture with the minimum number of sources. Experimentally, it was found that this square factor was important to improve the success rate of the method (more details can be found at [4, p. 148]). Once computed $S(\mathcal{C}_i)$ for all the combinations at $\mathcal{C}(t)$, the one with highest score is selected:

$$\hat{\mathcal{C}}(t) = \arg\max_i \{S(\mathcal{C}_i(t))\} \tag{10}$$

3 Extension using neighbor frames

In the previously described method, each frame was independently analysed, yielding the combination of fundamental frequencies that maximizes a given measure. One of the main limitations of this approach is that the window size (93 ms) is relatively short to perceive the pitches in a complex mixture, even for an expert musician. Context is very important in music to disambiguate certain situations. In this section the core method is extended, considering information about adjacent frames to produce a smoothed detection across time.

3.1 Temporal smoothing

A simple and effective novel technique is presented in order to smooth the detection across time. Instead of selecting the most likely combination at isolated frames, adjacent frames are also analysed to get the score of each combination.

The method aims to enforce the pitch continuity in time. For this, the fundamental frequencies of each combination $\mathcal{C}$ are mapped into music pitches, obtaining a pitch combination $\mathcal{C}'$. For instance, the combination $\mathcal{C}_i = \{261 \text{ Hz}, 416 \text{ Hz}\}$ is mapped into $\mathcal{C}'_i = \{\text{C}_4, \text{G}\sharp_4\}$.

If there is more than one combination with the same pitches (for instance, $\mathcal{C}_1 = \{260 \text{ Hz}\}$ and $\mathcal{C}_2 = \{263 \text{ Hz}\}$ are both $\mathcal{C}' = \{\text{C}_4\}$), it is removed, and the unique combination with the highest score value is only kept.

Then, at each frame t, a new smoothed score function $\tilde{S}(\mathcal{C}'_i(t))$ for a combination $\mathcal{C}'_i(t)$ is computed using the neighbor frames:

$$\tilde{S}(\mathcal{C}'_i(t)) = \sum_{j=t-K}^{t+K} S(\mathcal{C}'_i(j)) \tag{11}$$

where $\mathcal{C}'_i$ are the combinations that appear at least once in the $2K+1$ frames considered. Note that the score values for the same combination are summed in the $2K$ frames around t to obtain $\tilde{S}(\mathcal{C}'_i(t))$. An example of this procedure is displayed in Figure 5 for $K = 1$. If $\mathcal{C}_i$ is missing for any $t - K < j < t + K$, it does not contribute to the sum.

In this new situation, the pitch combination at the target frame t is selected as:

$$\hat{\mathcal{C}}'(t) = \arg\max_i \{\tilde{S}(\mathcal{C}'_i(t))\} \tag{12}$$

If $\hat{\mathcal{C}}'(t)$ does not contain any combination because there are no valid candidates in the frame t, then a rest is yielded without evaluating the adjacent frames.

This technique smoothes the detection in the temporal dimension. For a visual example, let's consider the smoothed intensity of a given candidate c' as:

$$\tilde{l}(c'(t)) = \sum_{j=t-K}^{t+K} l(c'(j)) \tag{13}$$

When the temporal evolution of the smoothed intensity $\tilde{l}(c'(t))$ of the winner combination candidates is plotted in a three-dimensional representation (see Figures 6 and 7), it can be seen that the correct estimates usually show smooth temporal curves. An abrupt change (a sudden note onset or offset, represented by a vertical cut in the smoothed intensities 3D plot) means that the energy of some harmonic components of a given candidate were suddenly improperly assigned to another candidate in the next frame. Therefore, vertical lines in the plot usually indicate errors in assigning harmonic components.

3.2 Pitch tracking

A basic pitch tracking method is introduced in order to favor smooth transitions of $\tilde{l}(c'(t))$. The proposed technique aims to increase the temporal coherence using a layered weighted directed acyclic graph (wDAG).
The idea is to minimize abrupt changes in the intensities of the pitch estimates. For that, a graph layered by frames is built with the pitch combinations, where the weights consider the differences in the smoothed intensities for the candidates in adjacent frames and their combination scores. Let $G = (V, v_I, E, w, t)$ be a layered wDAG, with vertex set V, initial vertex v_I, edge set E, and edge function w, where $w(v_i, v_j)$ is the weight of the edge from the vertex v_i to v_j. The position function $t : V \to \{0, 1, 2, \ldots, T\}$ associates each node with an input frame, being T the total number of frames. Each vertex $v_i \in V$ represents a combination $\mathcal{C}'_i$. The vertices are organized in layers (see Figure 8), in such a way that all vertices in a given layer have the

same value for $t(v) = \tau$, and they represent the M most likely combinations at a time frame τ.

The edges connect all the vertices of a layer with all the vertices of the next layer so, if $(v_i, v_j) \in E$, then $t(v_i) = \tau$ and $t(v_j) = \tau + 1$. The weights $w(v_i, v_j)$ between two combinations are computed as follows:

$$w(v_i, v_j) = \frac{D(v_i, v_j)}{\tilde{S}(v_j) + 1} \tag{14}$$

where $\tilde{S}(v_j)$ is the salience of the combination in v_j and $D(v_i, v_j)$ is a similarity measure for two combinations v_i and v_j, corresponding to the sum of the absolute differences between the intensities of all the candidates in both combinations:

$$D(v_i, v_j) = \sum_{\forall c \in v_i, v_j} |\tilde{l}(v_{i,c}) - \tilde{l}(v_{j,c})| \; + \sum_{\forall c \in v_i - v_j} \tilde{l}(v_{i,c}) \; + \sum_{\forall c \in v_j - v_i} \tilde{l}(v_{j,c}) \tag{15}$$

Using this scheme, the transition weight between two combinations considers the score of the target combination and the differences between the candidate intensities.

Once the graph is generated, the shortest path that minimizes the sum of weights from the starting node to the final state across the wDAG is found using the Dijkstra [33] algorithm. The vertices that belong to the shortest path are the pitch combinations yielded at each time frame.

Building the wDAG for all possible combinations at all frames could be computationally intractable, but considering only the M most likely combinations at each frame keeps almost the same runtime than without performing tracking for small values of M.

4 Evaluation

Initial experiments were done using a data set of random mixtures to perform a first evaluation and set up the parameters. Then, the proposed approaches were publicly evaluated and compared by a third party to other studies in the MIREX [22, 23] multiple f_0 estimation and tracking contest.

4.1 Evaluation metrics

Different metrics for multiple f_0 estimation can be found in the literature. The evaluation can be done both at frame by frame and note levels. The first mode evaluates the correct estimation in a frame by frame basis, whereas note tracking also considers the temporal coherence of the detection, adding more restrictions for a note to be considered correct. For instance, in the MIREX note tracking contest, a note is correct if its f_0 is closer than half a semitone to the ground-truth pitch and its onset is within a ± 50 ms range of the ground truth note onset.

A false positive (FP) is a detected pitch (or note, if evaluation is performed at note level) which is not present in the signal, and a false negative (FN) is a missing pitch. Correctly detected pitches (OK) are those estimates that are also present in the ground-truth at the detection time.

A commonly used metric for frame-based evaluation is the accuracy, defined as:

$$Acc = \frac{\Sigma_{OK}}{\Sigma_{OK} + \Sigma_{FP} + \Sigma_{FN}} \tag{16}$$

Alternatively, the performance can be assessed using precision/recall terms. Precision is related to the fidelity whereas recall is a measure of completeness.

$$\text{Prec} = \frac{\Sigma_{OK}}{\Sigma_{OK} + \Sigma_{FP}} \tag{17}$$

$$\text{Rec} = \frac{\Sigma_{OK}}{\Sigma_{OK} + \Sigma_{FN}} \tag{18}$$

The balance between precision and recall, or F-measure, is computed as their harmonic mean:

$$\text{F-measure} = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}} = \frac{\Sigma_{OK}}{\Sigma_{OK} + \frac{1}{2}\Sigma_{FP} + \frac{1}{2}\Sigma_{FN}} \tag{19}$$

An alternative metric based on the speaker diarization error score from NIST[a] was proposed by Poliner and Ellis [34] to evaluate multiple f_0 estimation methods. The NIST metric consists of a single error score which takes into account substitution errors (mislabeling an active voice, E_{subs}), miss errors (when a voice is truly active but results in no transcript, E_{miss}), and false alarm errors (when an active voice is reported without any underlying source, E_{fa}).

This metric avoids counting errors twice as classical metrics do in some situations. For instance, using accuracy, if there is a C_3 pitch in the reference ground-truth but the system reports a C_4, two errors (a false positive and a false negative) are counted. However, if no pitch was detected, only one error would be reported.

To compute the total error (E_{tot}) in T frames, the estimated pitches at every frame are denoted as N_{sys}, the ground-truth pitches as N_{ref}, and the number of correctly detected pitches as N_{corr}, which is the intersection between N_{sys} and N_{ref}.

$$E_{tot} = \frac{\sum_{t=1}^{T} \max\{N_{ref}(t), N_{sys}(t)\} - N_{corr}(t)}{\sum_{t=1}^{T} N_{ref}(t)} \tag{20}$$

Poliner and Ellis [34] state that, as in the universal practice in the speech recognition community, this is probably the most adequate measure, since it

gives a direct feel for the quantity of errors that will occur as a proportion of the total quantity of notes present.

4.2 Parameterization

A data set of random pitch combinations, also used in the evaluation of Klapuri [35] method, was used to tune up the algorithm parameters. The data set consists on 4000 mixtures with polyphony[b] 1, 2, 4, and 6. The 2842 audio samples from 32 music instruments used to generate the mixtures are from the McGill University master samples collection[c], the University of Iowa[d], IRCAM studio online[e], and recordings of an acoustic guitar. In order to respect the copyright restrictions, only the first 185 ms of each mixture[f] were used for evaluation. In this dataset, the range of valid pitches is [$f_{\min} = 38$ Hz, $f_{\max} = 2100$ Hz], and the maximum polyphony is $P = 6$.

The values for the free parameters of the method were experimentally evaluated. Their impact on the performance and efficiency can be seen on Figures 9 and 10, and it is extensively analysed in [4, pp. 141–156]. In these figures, the cross point represents the values selected for the parameters. Lines represent the impact tuning individual parameters when keeping the selected values for the rest of parameters.

In the parameterization stage, the selected parameter values were not those that achieved the highest accuracy in the test set, but those that obtained a good trade-off between accuracy and low computational cost.

The chosen parameter values for the core method are shown in Table 1. For the extended method, when considering K adjacent frames, different values for parameters $H = 15$, $\eta = 0.15$, $\kappa = 4$, and $\varepsilon = 0$ showed to perform slightly better, therefore they were selected for comparing the method to other studies (see Sec. 4.3). A detailed analysis of the parameterization process can be

found in [4].

4.3 Evaluation and comparison with other methods

The core method was externally evaluated and compared with other approaches in MIREX 2007 [22] multiple f_0 estimation and tracking contest, whereas the extended method was submitted to MIREX 2008 [23]. The data set used in both MIREX editions were essentially the same, therefore the results can be directly compared. The details of the evaluation and the ground-truth labeling are described in [36]. Accuracy, precision, recall and E_{tot} were reported for frame by frame estimation, whereas precision, recall and F-measure were used for the note tracking task.

The core method (PI1-07) was evaluated using the parameters specified in Table 1. For this contest, a final postprocessing stage was added. Once the fundamental frequencies were estimated, they were converted into music pitches, and pitch series shorter than $d = 56$ ms were removed to avoid some local discontinuities.

The extended method was submitted with pitch tracking (PI1-08) and without it (PI2-08) for comparison. In the non-tracking case, a similar procedure than in the core method was adopted, removing notes shorter than a minimum duration and merging note with short rests between them. Using pitch tracking, the methodology described in Sec. 3.2 was performed instead, increasing the temporal coherence of the estimate with the wDAG using $M = 5$ combinations at each layer.

The Table 2 shows all the methods evaluated. The proposed approaches were submitted both for frame by frame and note tracking contests, despite the only method which performs pitch tracking is PI1-08.

In the review from Bay et al. [36], the results of the algorithms evaluated in

both MIREX editions are analysed. As shown in Figure 11, the proposed methods achieved a high overall accuracy and the highest precision rates. The extended method also obtained the lowest error rates E_{tot} from all the methods submitted in both editions (see Figure 12).

In the evaluation of note tracking considering only onsets, the proposed approaches showed lower accuracies (Figure 13), as only the extended method can perform pitch tracking. The inclusion of the tracking stage did not improve the results for frame by frame estimation, but in the note tracking task the results outperformed those obtained for the same method without tracking. The proposed methods were also very efficient respect to the other state of the art algorithms presented (see Table 3), especially considering that they are based on a joint estimation scheme.

While the proposed approaches achieved the lowest E_{tot} score, there were very few false alarms compared to miss errors. On the other hand, the methods from Ryynänen and Klapuri [17] and Yeh et al. [37] had a better balanced precision, recall, as well as a good balance in the three error types, and as a result, high accuracies.

Quoting Bay et al. [36], "Inspecting the methods used and their performances, we can not make generalized claims as to what type of approach works best. In fact, statistical significance testing showed that the top three methods (YRC, PI, and RK) were not significantly different."

5 Conclusions and discussion

In this study, an efficient methodology is proposed for multiple f_0 estimation of real music signals assuming spectral smoothness and strong harmonic content without any other *a priori* knowledge of the sources.

The method can infer and evaluate hypothetical spectral patterns from the

analysis of different hypotheses taking into account the interactions with other sources.

The algorithm is extended considering adjacent frames to smooth the temporal detection. In order to increase the temporal coherence of the detection, a novel pitch tracking stage based on a wDAG has been included. The proposed algorithms were evaluated and compared to other works by a third party in a public contest (MIREX), obtaining a high accuracy, the highest precision and the lowest E_{tot} among all the multiple f_0 methods submitted. Although many possible combinations of candidates are evaluated at each frame, the presented approach has a very low computational cost, showing that it is possible to make an efficient joint estimation method by applying some constraints, like the sparse representation of only certain spectral peaks, the candidate filtering stage, and the combination pruning process.

The pitch tracking stage could be replaced by a more reliable method in a future study. For instance, the transition weights could be learned from a labeled test set, or a more complex tracking method like the high-order HMM scheme from Chang et al. [38] could be used instead. Besides intensity, the centroid of an HPS should also have a temporal coherence when belonging to the same source, therefore this feature could also be considered for tracking. Using stochastic models, a probability could be assigned to each pitch in order to remove those that are less probable given their context. Musical probabilities can be taken into account, like in [17], to remove very unlikely notes. The adaptation to polyphonic music of the stochastic approach from Perez-Sancho [39] is also planned as future study, in order to complement the multiple f_0 estimation method to obtain a musically coherent detection. Besides frame by frame analysis and the analysis of adjacent frames, the possibility of the extended method for combining similar information across

frames allows to consider different alternative architectures.

This novel methodology permits interesting schemes. For example, the beginnings of musical events can be estimated using an onset detection algorithm like [40]. Then, combinations of those frames that are between two consecutive onsets can be merged to yield the pitches within the inter-onset interval. This technique is close to segmentation, and it can obtain reliable results when the onsets are correctly estimated, as it happens with sharp attack sounds like piano, but a wrong estimate in the onset detection stage will affect the results.

Beats, that can be defined as a sense of equally spaced temporal units [41], can also be detected to merge combinations with a quantization grid. Once the beats are estimated (for example with a beat tracking algorithm like BeatRoot [42]), a grid split with a given beat divisor $1/q$ can be used, assuming that the minimum note duration is q. For instance, if $q = 4$, each inter-beat interval can be split in q sections. Then, the combinations of the frames that belong to the quantization unit can be merged to obtain the results at each minimum grid unit. Like in the onset detection scheme, the success rate of this approach depends on the success of the beat estimation.

The extended method can be applied using any of these schemes. The adequate choice of the architecture depends on the signal to be analysed. For instance, for timbres with sharp attacks, it is recommended to use onset information, which is very reliable for these kind of sounds. These alternative architectures have been perceptually evaluated using some example real songs, but a more rigorous evaluation of these schemes is left for future study, since an aligned dataset of real musical pieces with symbolic data is required for this task.

Competing interests

The authors declare that they have no competing interests.

Acknowledgment

This study was supported by the project DRIMS (code TIN2009-14247-C02), the Consolider Ingenio 2010 research programme (project MIPRCV, CSD2007-00018), and the PASCAL2 Network of Excellence, IST-2007-216886.

Endnotes

[a]National Institute of Standards and Technology.

[b]There are 1000 mixtures for each polyphony.

[c]http://www.music.mcgill.ca/resources/mums/html/index.htm

[d]http://theremin.music.uiowa.edu/MIS.html

[e]http://forumnet.ircam.fr/402.html?&L=1

[f]The authors would like to thank A. Klapuri for providing this data set for evaluation.

References

1. A Pertusa, JM Iñesta, Multiple fundamental frequency estimation using Gaussian smoothness, in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. I (Las Vegas, NV, 2008), pp. 105–108

2. A Klapuri, M Davy, *Signal Processing Methods for Music Transcription*. Springer Science+Business Media LCC, New York (2006)

3. MG Christensen, A Jakobsson, Multi-Pitch estimation, in *Synthesis Lectures on Speech and Audio Processing*, (Morgan & Claypool publishers, Seattle, WA, USA, 2009)

4. Pertusa A, Computationally efficient methods for polyphonic music transcription. *PhD thesis*, Universidad de Alicante (2010)

5. A Tolonen, M Karjalainen, A computationally efficient multipitch analysis model. IEEE Trans. Speech Audio Process. **8**(6), 708–716 (2000)

6. G Peeters, Music pitch representation by periodicity measures based on combined temporal and spectral representations, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. V (Toulouse, France, 2006) pp. 53–56

7. R Zhou, JD Reiss, M Mattavelli, G Zoia, A computationally efficient method for polyphonic pitch estimation. EURASIP J. Adv. Signal Process. **2009**(28) (2009)

8. M Marolt, Networks of adaptive oscillators for partial tracking and transcription of music recordings. J. New Music Res. **33**, 49–59 (2004)

9. A Pertusa, JM Iñesta, Polyphonic monotimbral music transcription using dynamic networks. Pattern Recogn. Lett. **26**(12), 1809–1818 (2005)

10. A Cont, Realtime multiple pitch observation using sparse non-negative constraints, in *Proc. of the 7th International Symposium on Music Information Retrieval (ISMIR)*, (Victoria, Canada, 2006), pp. 206–211

11. N Bertin, R Badeau, E Vincent, Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. IEEE Trans. Audio. Speech Language Process. **18**(3), 538–549 (2010)

12. JJ Carabias-Orti, P Vera-Candeas, FJ Cañadas-Quesada, N Ruiz-Reyes, Music scene-adaptive harmonic dictionary for unsupervised note-event detection. IEEE Trans. Audio Speech Language Process. **18**(3), 473–486 (2010)

13. P Leveau, E Vincent, G Richard, L Daudet, Instrument-specific harmonic atoms for mid-level music representation. IEEE Trans. Audio Speech Language Process. **16**, 116–128 (2008)

14. AT Cemgil, Bayesian Music Transcription. *PhD thesis*, Radboud University of Nijmegen, Netherlands (2004), http://www-sigproc.eng.cam.ac.uk/ atc27/papers/cemgil-thesis.pdf

15. M Davy, An introduction to signal processing. In *Signal Processing methods for music transcription*, (Springer Science+Business Media LCC, New York, 2006)

16. A Klapuri, Multipitch analysis of polyphonic music and speech signals using an auditory model. IEEE Trans. Audio, Speech and Language Process. **16**(2), 255–266 (2008)

17. M Ryynänen, A Klapuri, Polyphonic Music Transcription Using Note Event Modeling, in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, New York, USA, 2005) 319–322

18. C Cao, M Li, J Liu, Y Yan, Multiple F0 estimation in polyphonic music, in *Proc. of the 3rd Music Information Retrieval Evaluation eXchange (MIREX)*, (Vienna, Austria, 2007)

19. C Yeh, A Roebel, X Rodet, Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. IEEE Trans. Audio, Speech Language Process. **18**(6) (2010), pp. 1116–1126

20. V Emiya, R Badeau, B David, Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches, in *Proc. European Signal Processing Conference (EUSIPCO)*, (Rhodes, Greece, 2008)

21. FJ Cañadas-Quesada, P Vera-Candeas, N Ruiz-Reyes, JJ Carabias-Orti, Polyphonic transcription based on temporal evolution of spectral similarity of Gaussian Mixture Models, In *17th European Signal Processing Conference (EUSIPCO)*, (Glasgow, Scotland, 2009), pp. 10–14

22. MIREX, Music Information Retrieval Evaluation eXchange. Multiple fundamental frequency estimation and tracking contest (2007), http://www.music-ir.org/mirex/wiki/2007:Multiple_Fundamental_Frequency_Estimation_&_Tracking_Results

23. MIREX, Music Information Retrieval Evaluation eXchange. Multiple fundamental frequency estimation and tracking contest (2008), http://www.music-ir.org/mirex/wiki/2008:Multiple_Fundamental_Frequency_Estimation_&_Tracking_Results

24. C Yeh, Multiple F0 estimation for MIREX 2007, in *Proc. of the 3rd Music Information Retrieval Evaluation eXchange (MIREX)*, (Vienna, Austria, 2007)

25. A Pertusa, JM Iñesta, Multiple fundamental frequency estimation based on spectral pattern loudness and smoothness, in *Proc. of the 3rd Music Information Retrieval Evaluation eXchange (MIREX)*, (Vienna, Austria, 2007)

26. C Yeh, A Röbel, X Rodet, Multiple fundamental frequency estimation of polyphonic music signals, in *IEEE, Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. III (Philadelphia, PA, 2005), pp. 225–228

27. FJ Cañadas-Quesada, P Vera-Candeas, N Ruiz-Reyes, R Mata-Campos, JJ Carabias-Orti, Note-event detection in polyphonic musical signals based on Harmonic matching pursuit and spectral smoothness. J. New Music Res. **37**(3), 167–183 (2008)

28. A Klapuri, Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. IEEE Trans. Speech Audio Process. **11**(6), 804–816 (2003)

29. R Badeau, V Emiya, B David, Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. I (Taipei, Taiwan, 2009), pp. 3073–3076

30. C Yeh, Multiple fundamental frequency estimation of polyphonic recordings. *PhD thesis*, Université Paris VI - Pierre et Marie Curie (2008)

31. A Klapuri, Signal processing methods for the automatic transcription of music. *PhD thesis*, Tampere Univ. of Technology (2004)

32. RC Maher, Evaluation of a method for separating digitized duet signals. J. Audio Eng. Soc. **38**, 956–979 (1990)

33. EW Dijkstra, A note on two problems in connexion with graphs. *Numerische Mathematik* **1**, 269–271 (1959)

34. GE Poliner, DPW Ellis, A Discriminative Model for Polyphonic Piano Transcription. EURASIP J. Adv. Signal Process. **2007** (2007)

35. A Klapuri, Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes, in *Proc. of the Int. Conference on Music Information Retrieval (ISMIR)*, (Victoria, Canada, 2006), pp. 216–221

36. M Bay, AF Ehmann, JS Downie, Evaluation of multiple-F0 estimation and tracking systems, in *Proc. of the 10th International Conference on Music Information Retrieval (ISMIR)*, (Kobe, Japan, 2009), pp. 315–320

37. C Yeh, A Roebel, WC Chang, Multiple F0 estimation for MIREX 08, in *Proc. of the 4th Music Information Retrieval Evaluation eXchange (MIREX)*, (Philadelphia, PA, 2008)

38. WC Chang, AWY Su, C Yeh, A Roebel, X Rodet, Multiple-F0 tracking based on a high-order HMM model, in *Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx)*, (Espoo, Finland, 2008), pp. 379–386

39. C Pérez-Sancho, Stochastic Language Models for Music Information Retrieval. *PhD thesis*, Universidad de Alicante, Spain (2009)

40. A Pertusa, A Klapuri, JM Iñesta, Recognition of note onsets in digital music using semitone bands. Lecture Notes in Computer Science **3773**, 869–879 (2005)

41. S Handel, *Listening: An introduction to the perception of auditory events.* Bradford Books MIT Press, Cambridge (1989)

42. Dixon S: Onset detection revisited, in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, (Montreal, Canada, 2006), pp. 133–137

43. A Cont, Real-time transcription of music signals: MIREX 2007 submission description, in *Proc. of the 3rd Music Information Retrieval Evaluation eXchange (MIREX)*, (Vienna, Austria, 2007)

44. C Cao, M Li , Multiple F0 estimation in polyphonic music (MIREX 2008), in *Proc. of the 4th Music Information Retrieval Evaluation eXchange (MIREX)*, (Philadelphia, PA, 2008)

45. JL Durrieu, G Richard, B David, Singer melody extraction in polyphonic signals using source separation methods, in *Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. I (Las Vegas, NV, 2008), pp. 169–172

46. V Emiya, R Badeau, B David, Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches, in *Proc. of the 4th Music Information Retrieval Evaluation eXchange (MIREX)*, (Philadelphia, PA, 2008)

47. K Egashira, N Ono, S Sagayama, Sequential estimation of multiple fundamental frequencies through Harmonic-Temporal-Structured clustering, in *Proc. of the 4th Music Information Retrieval Evaluation eXchange (MIREX)*, (Philadelphia, PA, 2008)

48. H Kameoka, T Nishimoto, S Sagayama, A Multipitch analyser based on harmonic temporal structured clustering. IEEE Trans. Audio Speech Language Process. **5**(3), 982–994 (2007)

49. M Groble, Multiple fundamental frequency estimation, in *Proc. of the 4th Music Information Retrieval Evaluation eXchange (MIREX)*, (Philadelphia, PA, 2008)

50. T Lidy, A Rauber, A Pertusa, JM Iñesta, Improving genre classification by combination of audio and symbolic descriptors using a transcription system, in *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR)*, (Vienna, Austria, 2007), pp. 61–66

51. P Leveau, A multipitch detection algorithm using a sparse decomposition with instrument-specific harmonic atoms, in *Proc. of the 3rd Music Information Retrieval Evaluation eXchange (MIREX)*, (Vienna, Austria, 2007)

52. G Reis, F Fernandez, A Ferreira, Genetic algorithm approach to polyphonic music transcription for MIREX 2008, in *Proc. of the 4th Music Information Retrieval Evaluation eXchange (MIREX)*, (Philadelphia, PA, 2008)

53. SA Raczynski, N Ono, S Sagayama, Multipitch analisys with harmonic nonnegative matrix approximation, in *Proc. of the 8th Int. Conference on Music Information Retrieval (ISMIR)*, (Vienna, Austria, 2007), pp. 381–386

54. E Vincent, N Bertin, R Badeau, Two nonnegative matrix factorization methods for polyphonic pitch transcription, in *Proc. of the 3rd Music Information Retrieval Evaluation eXchange (MIREX)*, (Vienna, Austria, 2007)

55. R Zhou, JD Reiss, A real-time polyphonic music transcription system, in *Proc. of the 4th Music Information Retrieval Evaluation eXchange (MIREX)*, (Philadelphia, PA, 2008)

Figure 1. General overview of the core method and its extension.

Figure 2. Partial selection example. The selected peak is the one with the greatest weighted value.

Figure 3. HPS estimation in a combination of two candidates separated by one octave. The HPS of f_1 is estimated by interpolation using the non-overlapped partials.

Figure 4. Spectral smoothness measure example. The normalized HPS vector $\bar{\mathbf{p}}$ and the smooth version $\tilde{\mathbf{p}}$ of two candidates c_1 (top) and c_2 (down) are shown. In this example, $r(c_1) = 0.13$, and $r(c_2) = 1.23$.

Figure 5. Example of combinations fusion across adjacent frames using $K = 1$.

Figure 6. 3D intensity representation (oboe). Top: Ground-truth evolution of pitch along time for an oboe melody. Bottom: 3D temporal representation of $\tilde{l}(c'(t))$ for the candidates of the winner combination at each frame. In this example, all the pitches were correctly detected.

Figure 7. 3D intensity representation (piano). Top: Ground-truth evolution of pitch along time for a piano piece. Bottom: 3D temporal representation of $\tilde{l}(c'(t))$ for the candidates of the winner combination at each frame. Most errors occur when there exist steep intensity transitions which mean that harmonics of a candidate were wrongly assigned to another candidate

Figure 8. Layered wDAG example for $M = 3$ combinations at each time. Each layer $t(v) = \tau$ represents a time frame, and each vertex is a combination $\mathcal{C}'_i(t)$. Weights have been multiplied by 10^4 for visual clarity. The grayed nodes are the pitch combinations selected at each frame in this example.

Figure 9. Accuracy for the core method in the random pitch dataset adjusting the individual parameters. The abcissae axis is not labeled since these values depend on each particular parameter (the first and last values for each parameter have been displayed in each plot to get the grid step for each parameter).

Figure 10. Core method runtime adjusting the parameters. Frame by frame method runtime in seconds for the entire random mixtures database for the parameters that have some influence in the computational cost.

Figure 11. Frame by frame MIREX accuracy. Figure from [36]. Frame by frame precision, recall and accuracy for MIREX 07-08 multiple f_0 estimation methods.

Figure 12. Frame by frame MIREX E_{tot}. Figure from [36], showing E_{subs}, E_{miss} and E_{fa} for MIREX 07-08 frame by frame evaluation ordered by E_{tot}.

Figure 13. MIREX note tracking F-m (Figure from [36]). Figure from [36]. Precision, recall, average F-measure and average overlap based on note onset for MIREX 07-08 note tracking.

Table 1. Parameter values experimentally selected

Stage	Parameter	Symbol	Value
Preprocessing	Partial selection threshold	μ	0.1
	Zero padding factor	z	4
Candidate selection	Min. f_0 amplitude	ε	2
Combination generation	Max. number of candidates	F	10
	Partial search bandwidth	f_r	11 Hz
	HPS size	H	10
Combination evaluation	Absolute intensity threshold	γ	5
	Relative intensity threshold	η	0.1
	Smoothness weight	κ	2
Temporal smoothing	Number of adjacent frames	K	2

ALE Nodes is the fastest machine. Runtime details are in [22, 23]

Table 2. MIREX 07-08 methods submitted for frame by frame (FBF) and note tracking (NT) evaluation.

Id	References	FBF	NT	Methodology
AC-07	[43]	√	√	Unsupervised learning
CL-07	[18]	√		Iterative cancelation
CL-08	[44]	√		Iterative cancelation
YRC-07	[30]	√		Joint estimation
DRD-08	[45]	√		Iterative cancelation
EBD-07	[20]	√	√	Statistical inference
EBD-08	[46]	√	√	Statistical inference
EOS-08	[47]	√	√	Statistical inference
EOS-07	[48]	√	√	Statistical inference
MG-08	[49]	√		Database matching
PE-07	[34]	√	√	Supervised learning
PI1-07	**[1]**	√	√	**Core method**
PI2-07	[50]		√	Iterative cancellation
PI1-08		√	√	**Extended method + tracking**
PI2-08		√	√	**Extended method**
PL-07	[51]	√		Matching pursuit
RFF-08	[52]	√	√	Supervised learning
RK-07	[17]	√	√	Iterative cancellation + f_0 tracking

Table 2. continued

Id	References	FBF	NT	Methodology
RK-08	[17]	√	√	Iterative cancellation + f_0 tracking
SR-07	[53]	√		Unsupervised learning
VBB-07	[54]	√	√	Unsupervised learning
VBB-08	[54]	√	√	Unsupervised learning
YRC1-08	[37]	√		Joint estimation
YRC2-08	[37]	√	√	Joint estimation + f_0 tracking
ZR-07	[7]	√		Signal processing
ZR-08	[55]		√	Signal processing

The presented methods are indicated in bold.

Table 3. MIREX 07-08 frame by frame and note tracking runtimes in seconds of the top ten accuracy methods

Id	FBF (s)	NT (s)	Machine
ZR-07	271		BLACK
PI1-07	364	364	ALE Nodes
PI2-08	792	790	ALE Nodes
PI1-08	955	950	ALE Nodes
VBB2-07	2233		ALE Nodes
RK-07	3540	3285	SANDBOX
RK-08	5058	5044	ALE Nodes
YRC1-08	57483		ALE Nodes
YRC2-08	57483	57483	ALE Nodes
YRC-07	132300		ALE Nodes

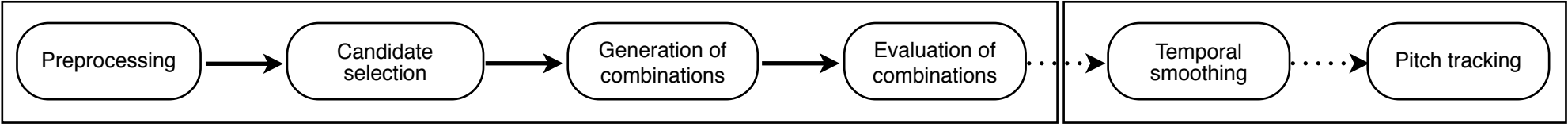

Preprocessing
Candidate selection
Generation of combinations
Evaluation of combinations
Temporal smoothing
Pitch tracking
Core method
Extended method

Figure 1

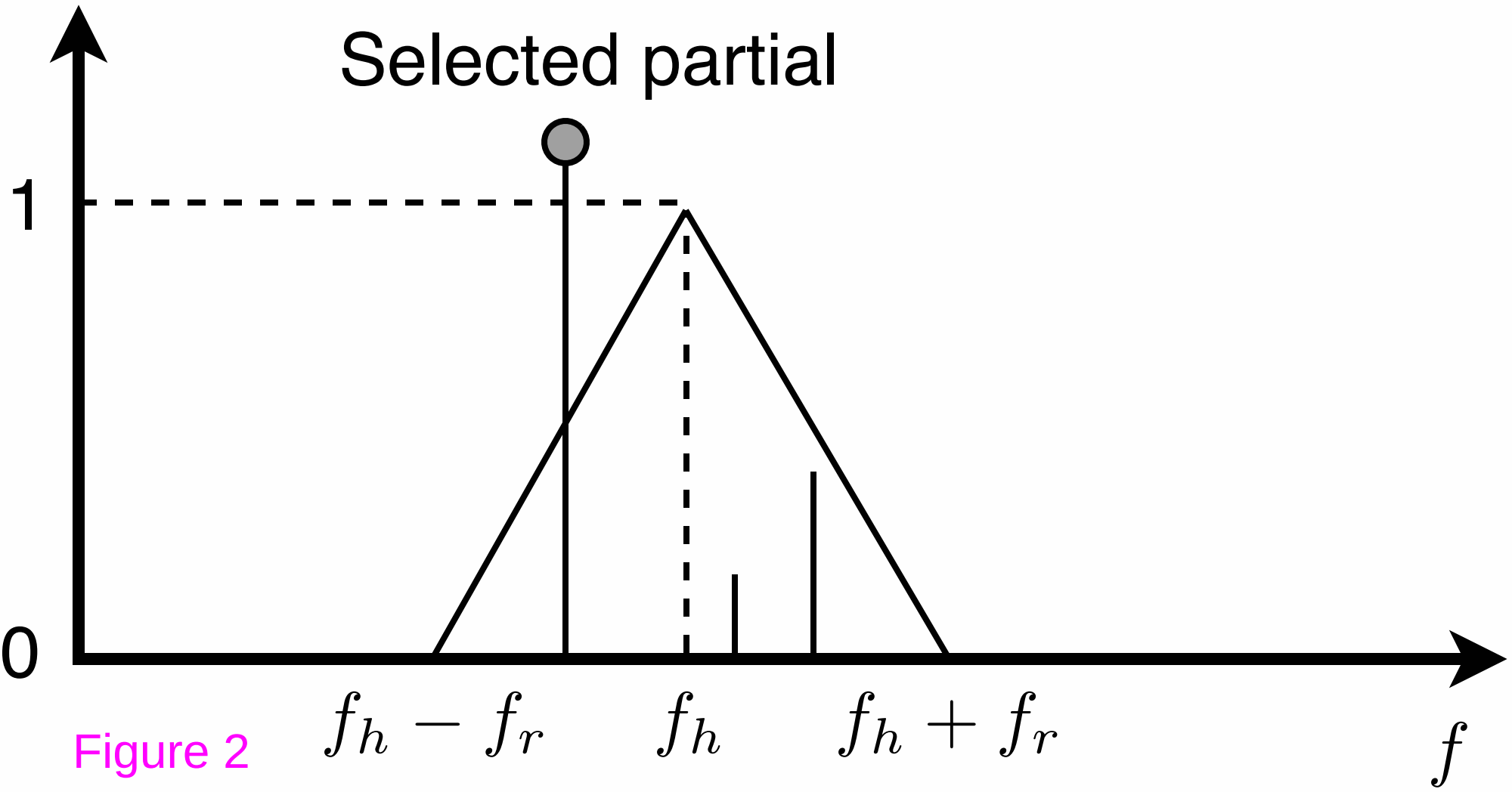


Figure 2

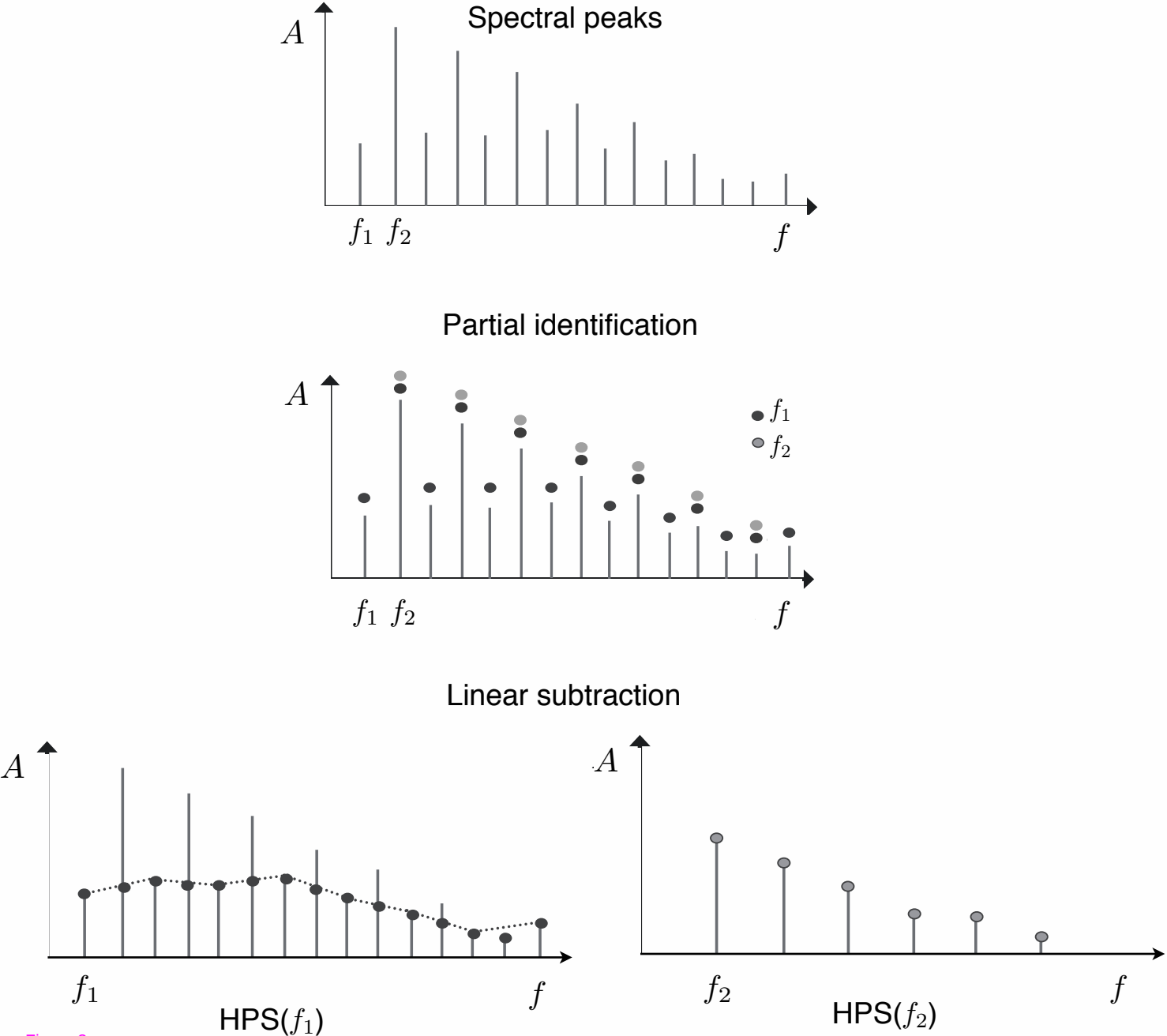
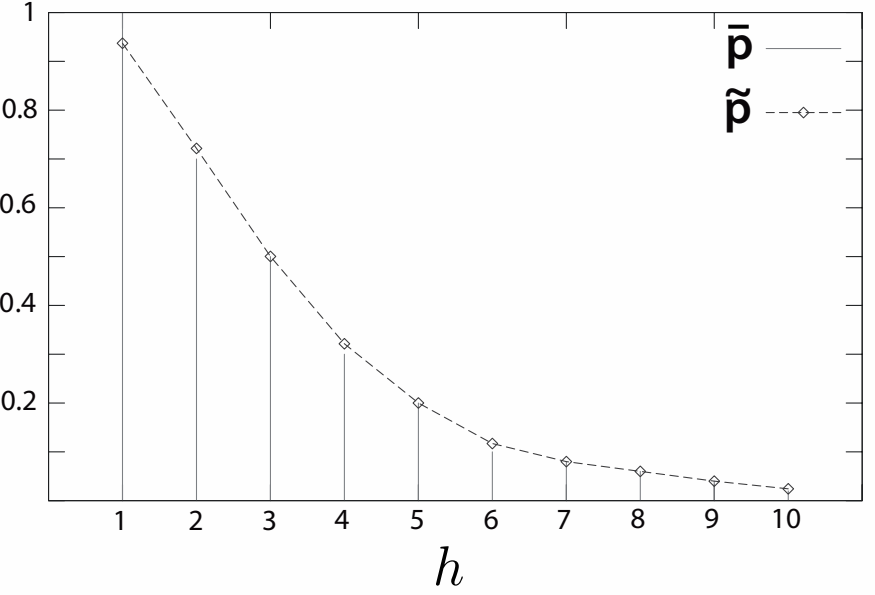

Spectral peaks
A
f_1 f_2
f
Partial identification
A
f_1
f_2
f_1 f_2
f
Linear subtraction
A
f_1
f
HPS(f_1)
A
f_2
f
HPS(f_2)

Figure 3



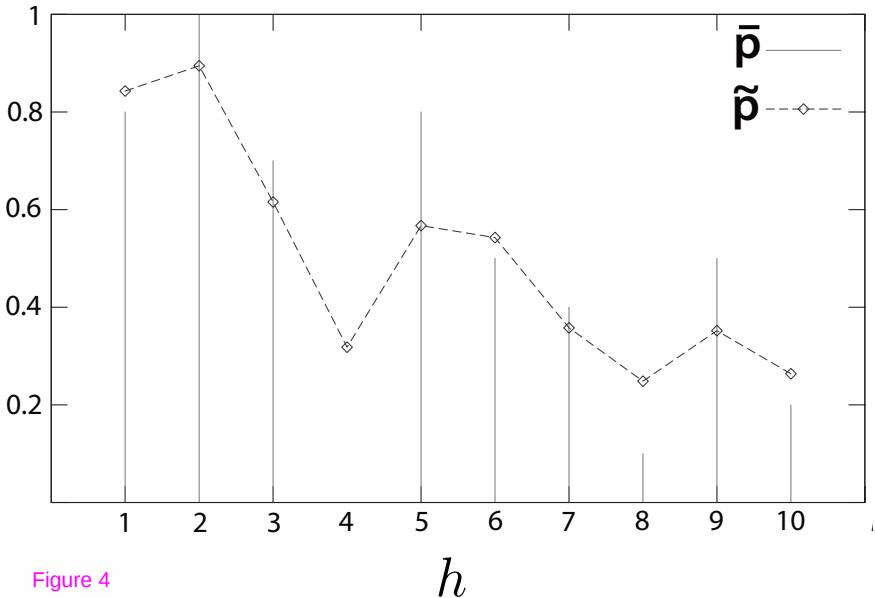


Figure 4

Combinations at $t-1$, t and $t+1$. Most likely combination $\hat{\mathcal{C}}(t)$ using the core method is hightlighted

$\mathcal{C}'_1(t-1)$ ={C3}	$S(\mathcal{C}'_1(t-1))$ =2100
$\mathcal{C}'_2(t-1)$ ={G3, E3}	$S(\mathcal{C}'_2(t-1))$ =1000
$\mathcal{C}'_3(t-1)$ ={E3, G4}	$S(\mathcal{C}'_3(t-1))$ =140

$\mathcal{C}'_1(t)$ ={C3, G4}	$S(\mathcal{C}'_1(t))$ =2000
$\mathcal{C}'_2(t)$ ={C3}	$S(\mathcal{C}'_2(t))$ =1800
$\mathcal{C}'_3(t)$ ={E3, G4}	$S(\mathcal{C}'_3(t))$ =200

$\mathcal{C}'_1(t+1)$ ={C3}	$S(\mathcal{C}'_1(t+1))$ =1700
$\mathcal{C}'_2(t+1)$ ={C3, G4}	$S(\mathcal{C}'_2(t+1))$ =1200
$\mathcal{C}'_2(t+1)$ ={E3}	$S(\mathcal{C}'_3(t+1))$ =100

Combination selection with temporal smoothing

$\mathcal{C}'_1(t)$ ={C3}	$\tilde{S}(\mathcal{C}'_1(t))$ =7400
$\mathcal{C}'_2(t)$ ={C3, G4}	$\tilde{S}(\mathcal{C}'_2(t))$ =3200
$\mathcal{C}'_3(t)$ ={G3, E3}	$\tilde{S}(\mathcal{C}'_3(t))$ =1000
$\mathcal{C}'_4(t)$ ={E3, G4}	$\tilde{S}(\mathcal{C}'_4(t))$ =340
$\mathcal{C}'_5(t)$ ={E3}	$\tilde{S}(\mathcal{C}'_5(t))$ =100

Figure 5

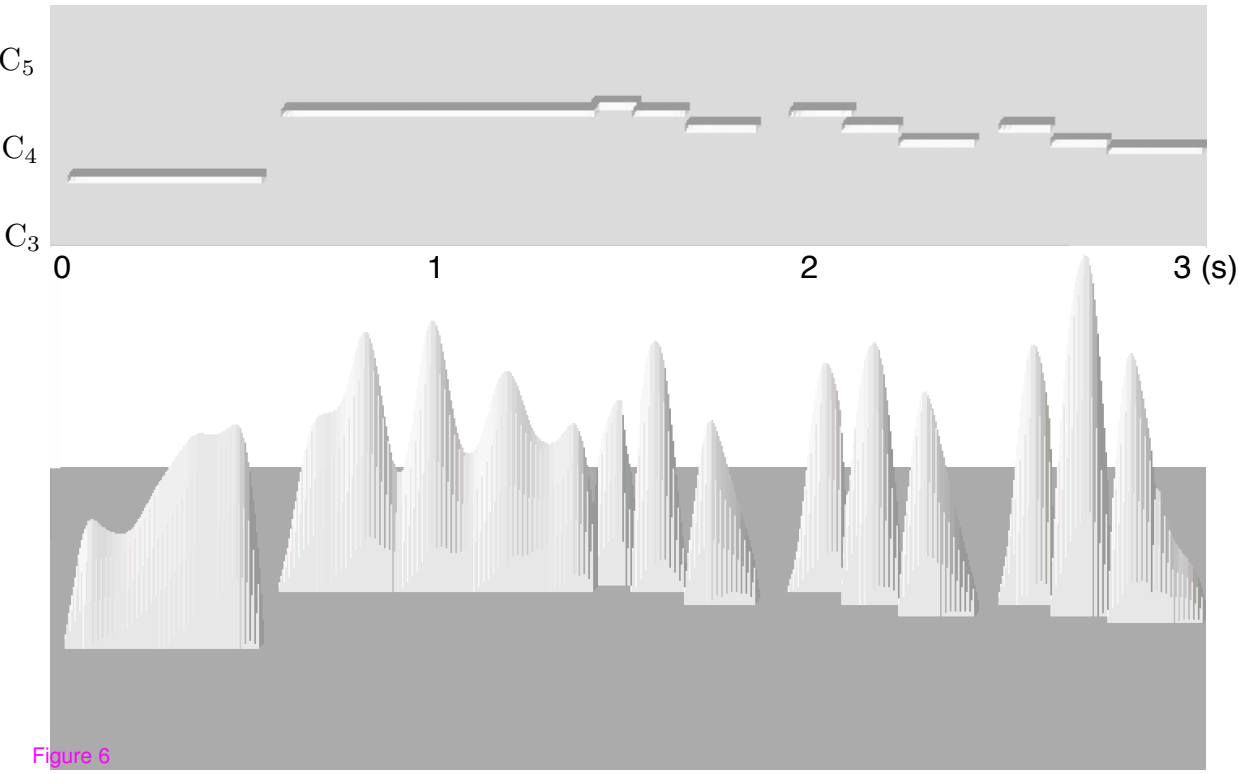


Figure 6

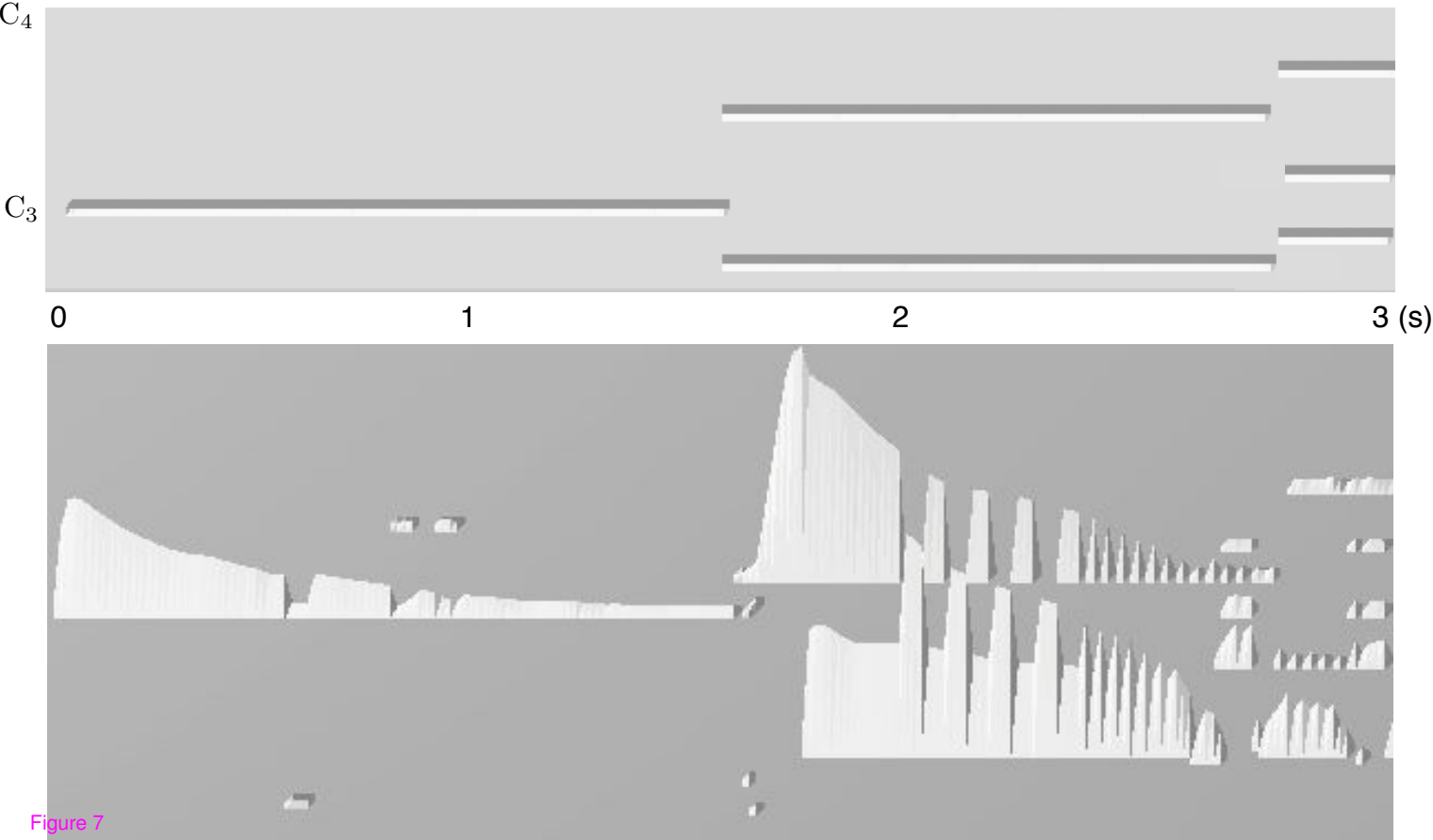


Figure 7

Figure 8

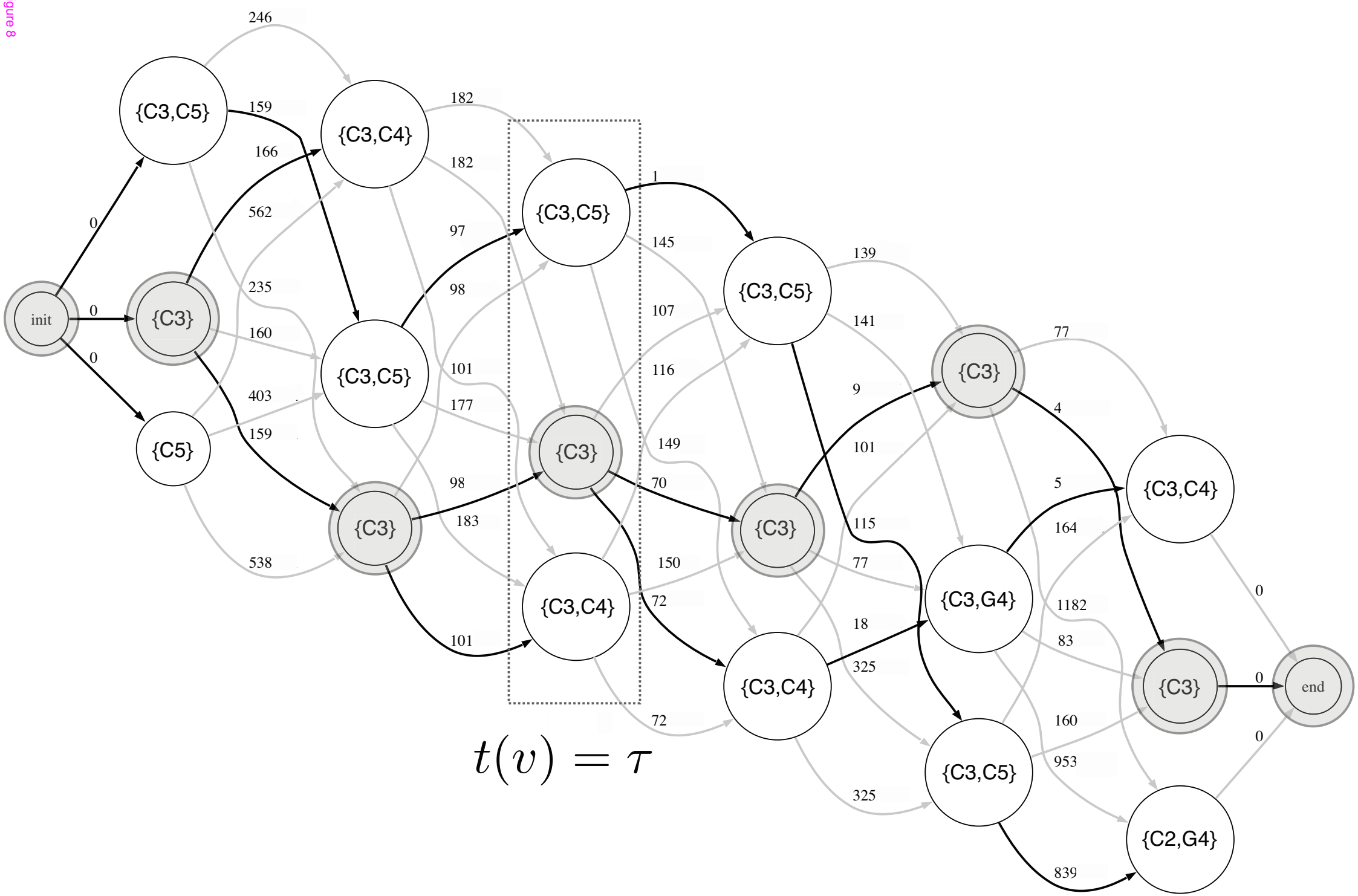
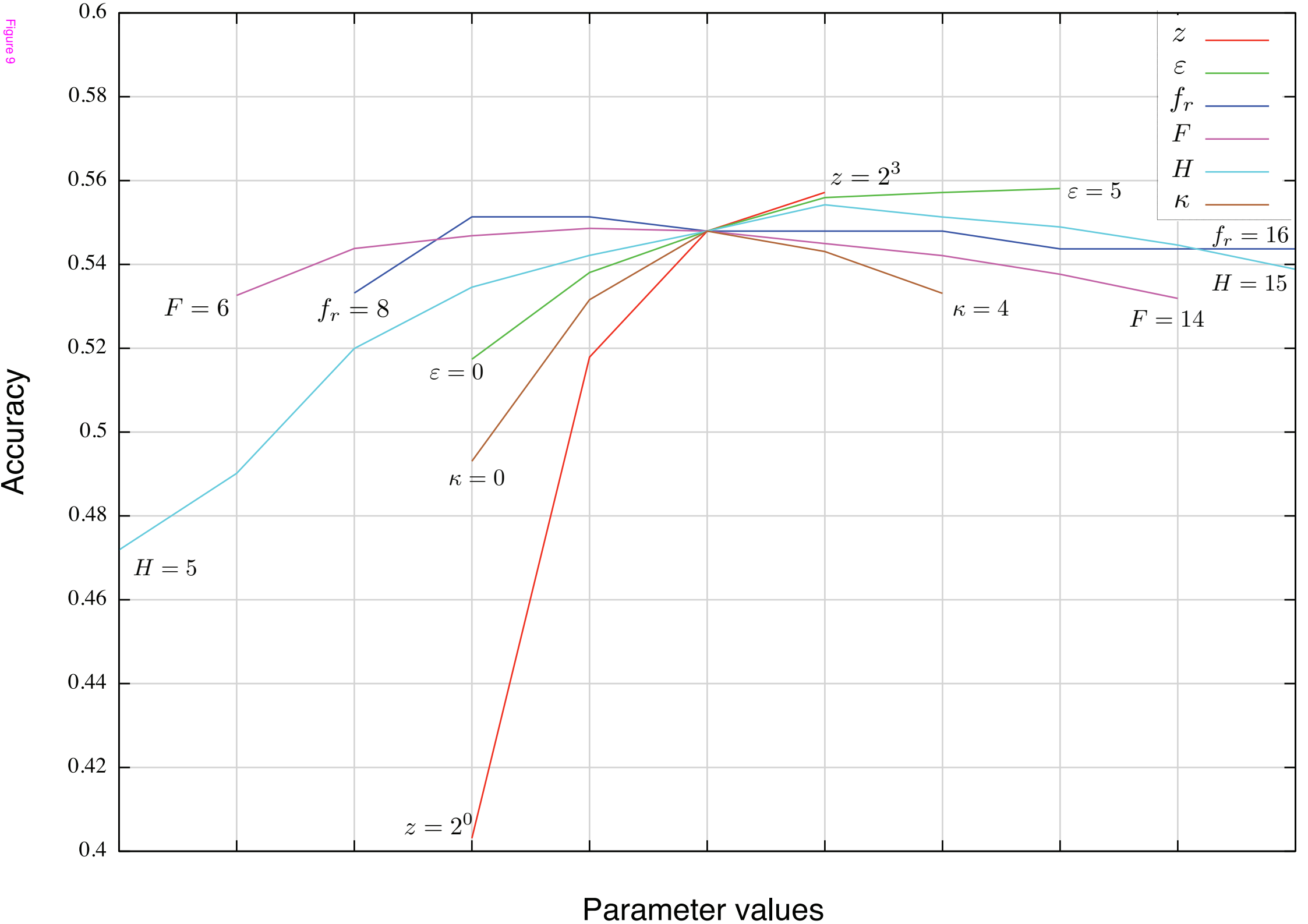

init
{C3}
{C5}
{C3,C5}
{C3,C4}
{C3,C5}
{C3}
{C3,C4}
{C3,C5}
{C3}
{C3,G4}
{C3,C5}
{C3}
{C3,C4}
{C2,G4}
end
246
159
166
562
235
160
403
159
538
182
182
97
98
101
177
98
183
101
1
145
107
116
149
70
150
72
72
139
141
9
101
115
77
18
325
325
77
4
5
164
1182
83
160
953
839
0
0
0
0
0
0
0
$t(v) = \tau$

Figure 9


Accuracy
Parameter values
0.6
0.58
0.56
0.54
0.52
0.5
0.48
0.46
0.44
0.42
0.4
z
ε
f_r
F
H
κ
$z = 2^3$
$\varepsilon = 5$
$f_r = 16$
$H = 15$
$F = 14$
$\kappa = 4$
$F = 6$
$f_r = 8$
$\varepsilon = 0$
$\kappa = 0$
$H = 5$
$z = 2^0$

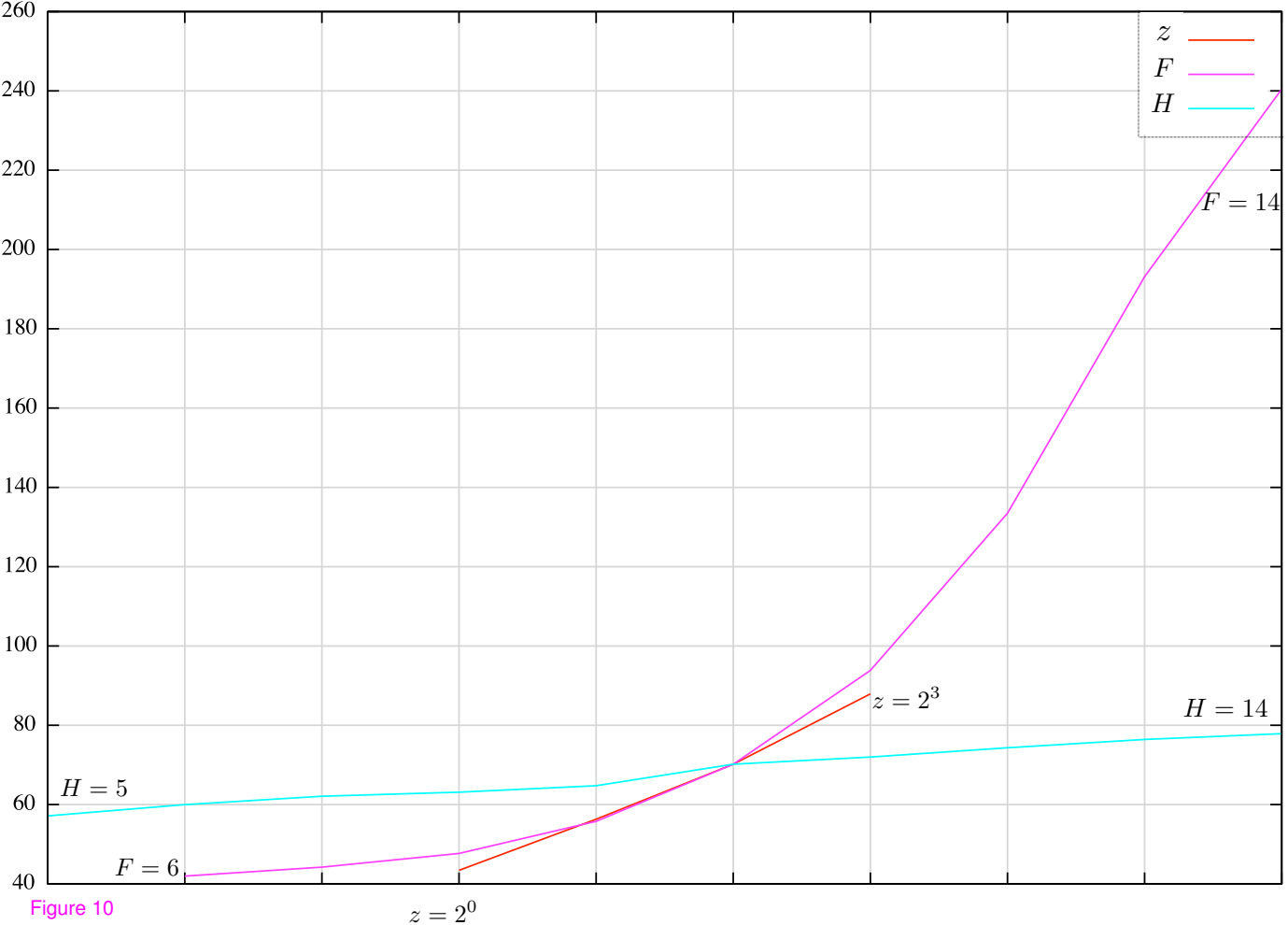

260
240
220
200
180
160
140
120
100
80
60
40
z
F
H
$F = 14$
$z = 2^3$
$H = 14$
$H = 5$
$F = 6$
$z = 2^0$

Figure 10

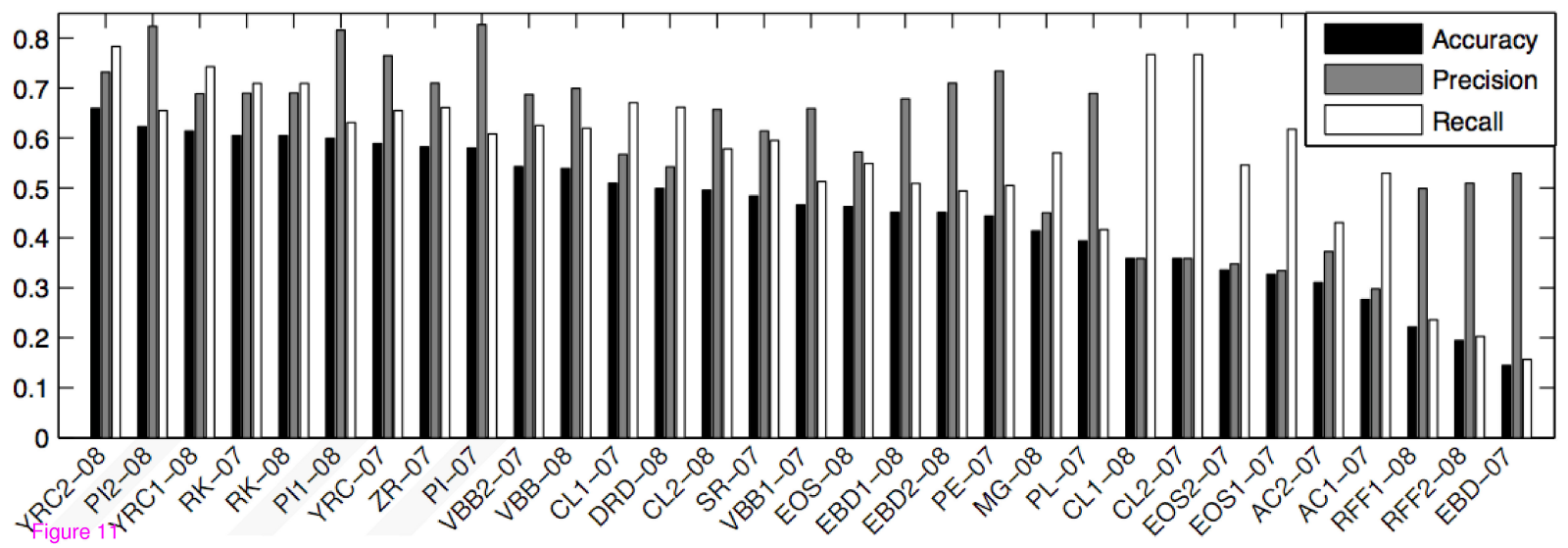

Accuracy
Precision
Recall
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
YRC2–08
PI2–08
YRC1–08
RK–07
RK–08
PI1–08
YRC–07
ZR–07
PI–07
VBB2–07
VBB–08
CL1–07
DRD–08
CL2–08
SR–07
VBB1–07
EOS–08
EBD1–08
EBD2–08
PE–07
MG–08
PL–07
CL1–08
CL2–07
EOS2–07
EOS1–07
AC2–07
AC1–07
RFF1–08
RFF2–08
EBD–07

Figure 11

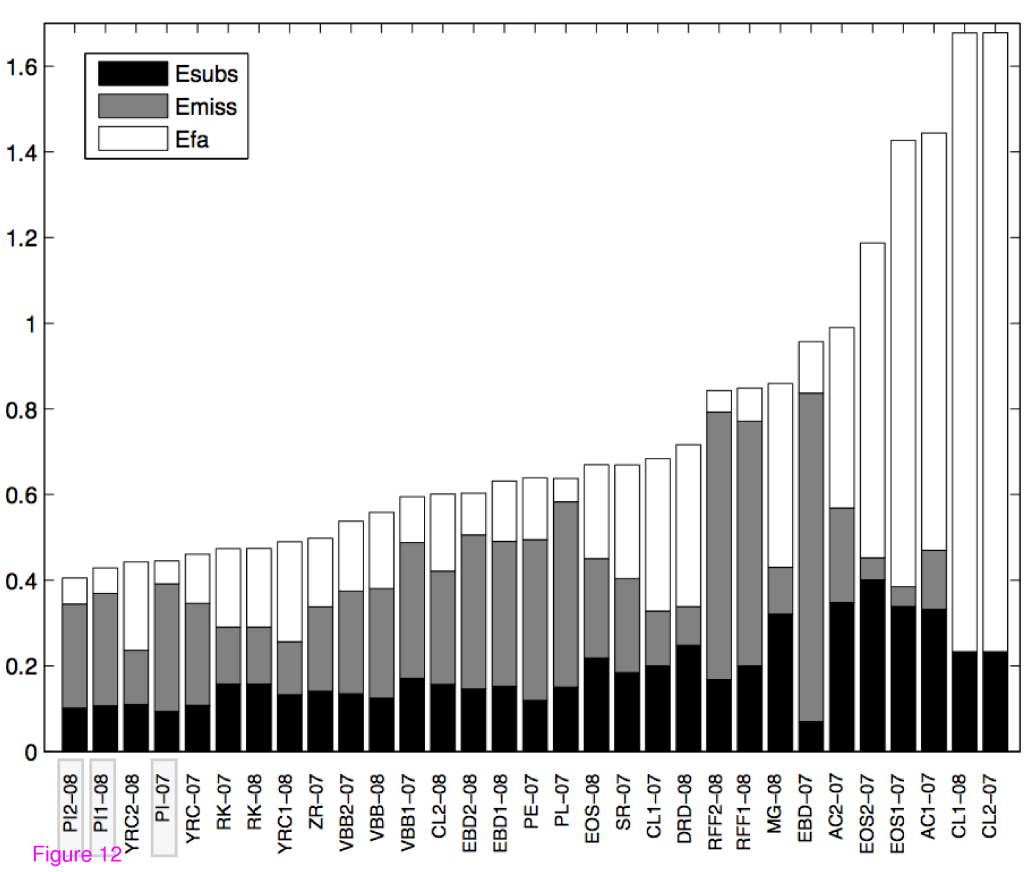

Esubs
Emiss
Efa
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
PI2−08
PI1−08
YRC2−08
PI−07
YRC−07
RK−07
RK−08
YRC1−08
ZR−07
VBB2−07
VBB−08
VBB1−07
CL2−08
EBD2−08
EBD1−08
PE−07
PL−07
EOS−08
SR−07
CL1−07
DRD−08
RFF2−08
RFF1−08
MG−08
EBD−07
AC2−07
EOS2−07
EOS1−07
AC1−07
CL1−08
CL2−07

Figure 12

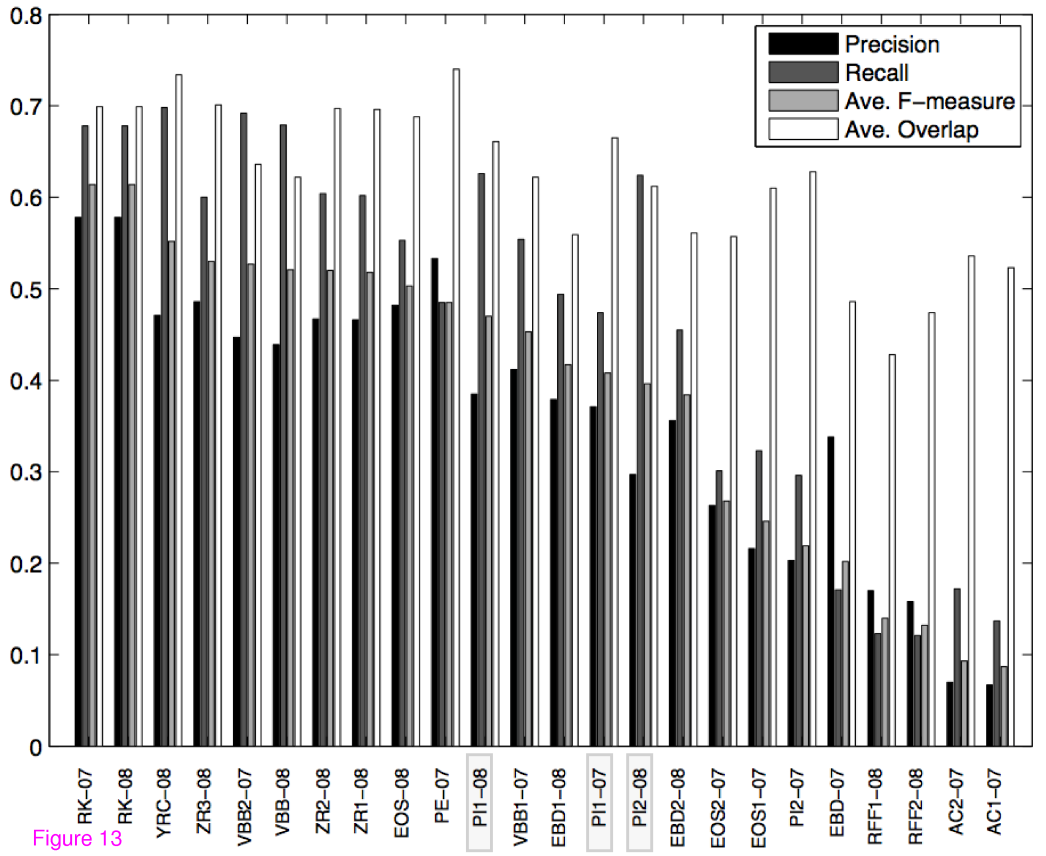

Precision
Recall
Ave. F−measure
Ave. Overlap
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
RK−07
RK−08
YRC−08
ZR3−08
VBB2−07
VBB−08
ZR2−08
ZR1−08
EOS−08
PE−07
PI1−08
VBB1−07
EBD1−08
PI1−07
PI2−08
EBD2−08
EOS2−07
EOS1−07
PI2−07
EBD−07
RFF1−08
RFF2−08
AC2−07
AC1−07

Figure 13