A CLASSIFICATION APPROACH TO MULTIPITCH ANALYSIS

Anssi Klapuri
Department of Signal Processing, Tampere University of Technology
klap@cs.tut.fi

ABSTRACT

This paper proposes a pattern classification approach to detecting the pitches of multiple simultaneous sounds. In order to deal with the octave ambiguity in pitch estimation, a statistical classifier is trained which observes the value of a detection function both at the position of a candidate pitch period and at its integer multiples and submultiples, in order to decide whether the candidate period should be accepted or rejected. The method improved significantly over a reference method in simulations.

1 INTRODUCTION

A fundamental problem of basically all pitch detection functions (such as the autocorrelation function) is that they do not show a peak only at the position of the true pitch, but also at twice and half the correct pitch, and often at all multiples and submultiples of it. This ambiguity is particularly challenging in multipitch detection where the detection function easily becomes congested with spurious peaks due to the ambiguity associated with each component sound.

To tackle the problem, multipitch estimation methods typically search for a set of pitch frequencies that best explain all the peaks in the detection function. Both joint estimation of multiple pitches and iterative detection and cancellation have been proposed (see [1, 2] for a review). A limitation of many of these techniques is that they produce a discrete set of detected pitch values, not a continuous function which would show the likelihoods of all pitch candidates within a given range. The latter would be more desirable for feature extraction purposes, where the actual detection stage is postponed to processes that look at a larger time scale and may include musicological constraints.

In this paper, we investigate a classification approach to pitch analysis. This approach has been previously investigated by Ellis and Poliner in [3], but they considered the multipitch analysis for a specific instrument (piano) and the applied technique was different from the present one.

SMC 2009, July 23-25, Porto, Portugal

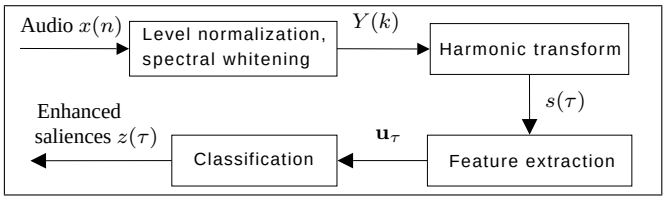

Figure 1. Overview of the method. See text for details.

2 METHOD

Figure 1 shows an overview of the proposed method. The first two steps, spectral whitening and harmonic transform, can be seen as preprocessing to distill information that is relevant for pitch detection. The steps are similar to the front-end used in [4] and produce a pitch salience function $s(\tau)$ where peaks indicate potential pitch periods in the input.

The latter two steps, feature extraction and classification, constitute the core of the method proposed here. They produce an enhanced salience function $z(\tau)$, where the peaks that correspond to correct periods are emphasized and extraneous ones are suppressed.

2.1 Level normalization and spectral whitening

The input audio signal $x(n)$ is blocked into 93 ms analysis frames that are processed independently. The signal within each frame is Hamming windowed, level-normalized to unity variance, zero-padded to twice its length, and then discrete Fourier transformed to obtain spectrum $X(k)$.

Spectral whitening, or flattening, is applied on $X(k)$ in order to suppress timbral information and thereby make the subsequent pitch analysis more robust to various sound sources. This is achieved by calculating power σ_c^2 of the signal within critical-band subbands c and by scaling the signal within each band by $\gamma_c = \sigma_c^{\nu-1}$, where $\nu = 0.16$ is a parameter determining the amount of whitening. The resulting whitened magnitude spectrum is denoted by $Y(k)$.

2.2 Harmonic transform

A harmonic transform is applied on the spectrum $Y(k)$ in order to calculate the saliences $s(\tau)$ of pitch period candi-

dates τ:

$$s(\tau) = \sum_{h=1}^{H} g(\tau, h) \max_{k \in \kappa_{\tau,h}} Y(k), \quad (1)$$

where the set $\kappa_{\tau,h}$ defines a range of frequency bins in the vicinity of the h:th overtone partial of the pitch candidate f_s/τ (f_s denoting the sampling rate) and $H = 20$. More exactly, $\kappa_{\tau,h} = \{\lfloor hK/(\tau + \Delta\tau/2) \rceil, \ldots, \lfloor hK/(\tau - \Delta\tau/2) \rceil\}$, where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, K is the length of the Fourier transform, and $\Delta\tau = 0.5$ denotes the spacing between successive period candidates τ.

The weights $g(\tau, h)$ are defined after [4] and are of the form $g(\tau, h) = (f_s/\tau + \epsilon_1)/(hf_s/\tau + \epsilon_2)$, where $\epsilon_1 = 52$ Hz and $\epsilon_2 = 320$ Hz. Note that the weights reduce to $1/h$ if the moderation terms ϵ_1 and ϵ_2 are omitted.

2.3 Feature extraction

Peaks in the salience function $s(\tau)$ are useful for indicating potential fundamental frequencies in the input signal. However, a pitched sound in the input does not only produce a peak at the corresponding pitch period τ, but also at multiples and submultiples of τ, complicating the pitch detection.

In order to do the detection more robustly, we observe $s(\tau)$ at the candidate period τ, but also at its multiples and submultiples. Let us define a vector $\mathbf{a}_\tau$:

$$\mathbf{a}_\tau = [1, s(\tau), s(2\tau), \ldots, s(J\tau), s(\tau/2), s(\tau/3), \ldots, s(\tau/J)]^\mathsf{T}$$

where $J = 5$ is the maximum (sub)multiple of τ considered. The length of $\mathbf{a}_\tau$ is $2J$. We then form a feature vector

$$\mathbf{v}_\tau = \begin{bmatrix} b_0(\tau)\mathbf{a}_\tau \\ b_1(\tau)\mathbf{a}_\tau \\ \vdots \\ b_{M-1}(\tau)\mathbf{a}_\tau \end{bmatrix}$$

where $b_m(\tau) = [\log(\tau + 1)]^m$, $m = 1, \ldots, M$, are basis functions that depend on the period τ and allow the subsequent statistical model to treat short and long periods differently. The length of $\mathbf{v}_\tau$ is $2JM$.

From here on, we consider data from different analysis frames and use $\mathbf{v}_{i,\tau}$ to denote the feature vector corresponding to period candidate τ in frame i. For the purpose of training, we collect $\mathbf{v}_{i,\tau}$ corresponding to the true periods in each frame, plus those corresponding to the 20 next-highest "false" peaks in $s(\tau)$. The vectors $\mathbf{v}_{i,\tau}$ in different frames and for different τ are stored as columns in a large matrix $\mathbf{V}$. The matrix is then processed by removing the uppermost row which is $b_0(\tau) \cdot 1 \equiv 1$ at all columns. The rest of the rows are normalized to zero mean and unity variance. The resulting normalized matrix is denoted by $\mathbf{W}$. The columns of $\mathbf{W}$ correspond to individual feature vectors, $\mathbf{w}_{i,\tau}$.

Finally, a linear transform is employed to decorrelate the features and to reduce their dimensionality. We tested principal component analysis (PCA) and linear discriminant analysis (LDA) for this purpose. They both produce a transform matrix $\mathbf{A}$ of size $((2JM-1) \times D)$. The transformed feature vectors $\mathbf{u}_{i,\tau}$ with dimensionality D are obtained by

$$\mathbf{u}_{i,\tau} = \mathbf{A}^\mathsf{T} \mathbf{w}_{i,\tau}. \quad (2)$$

2.4 Classification

Gaussian mixture models (GMMs) are used to classify the peaks in $s(\tau)$ either as "true" or "false" pitch periods. A GMM is defined as

$$\mathsf{p}(\mathbf{u}_{i,\tau}|\theta) = \sum_{j=1}^{J} \beta_j \mathcal{N}(\mathbf{u}_{i,\tau}; \mu_j, \mathbf{\Sigma}_j), \quad (3)$$

where $\mathcal{N}(\mathbf{u}; \mu, \mathbf{\Sigma})$ denotes Gaussian distribution with mean μ and covariance $\mathbf{\Sigma}$. The shorthand $\theta = \{\beta_j, \mu_j, \mathbf{\Sigma}_j\}$ is used to refer to all the parameters of a GMM.

Two GMM models are trained, using the feature vectors corresponding to the "true" and "false" periods, respectively. The resulting model parameters are denoted by θ_T and θ_F, respectively.

Enhanced salience $z_i(\tau)$ of period candidate τ in frame i is then defined as

$$z_i(\tau) = \log \mathsf{p}(\mathbf{u}_{i,\tau}|\theta_\mathrm{T}) - \log \mathsf{p}(\mathbf{u}_{i,\tau}|\theta_\mathrm{F}). \quad (4)$$

The above formula calculates salience as the difference of the log-likelihoods for the two models. It is important to use the model for the false peaks as a "background" model in (4): including only the first term on the right-hand side of (4) would give a low salience for an exceptionally strong peak since it does not fit ideally to the model of true peaks. Calculating the salience as the difference between the two models corrects this problem, since these exceptionally strong cases are even less likely in the background model.

3 RESULTS

The proposed method was tested on mixtures of 1, 2, 4, and 6 simultaneous sounds, randomly mixing sounds from 32 different musical instruments. Half of the data was contaminated with random drum sounds using 0 dB SNR. The models were trained using 2600 sound mixtures and tested using a set of 1300 different mixtures. The results are averaged over the test cases.

Instrument samples were obtained from the McGill University Master Samples collection, the University of Iowa website, IRCAM Studio Online, and by making independent recordings for the acoustic guitar. Instruments represented are the piano, the guitar, mallet percussions (marimba, vibraphone), brass and reed instruments, strings, and flutes.

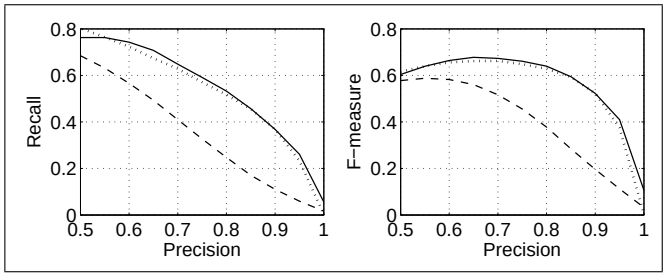

Figure 2. The left panel shows precision and recall for the proposed method with LDA (solid line), with PCA (dotted line) and for the baseline method (dashed line). The right panel shows F-measure as a function of precision.

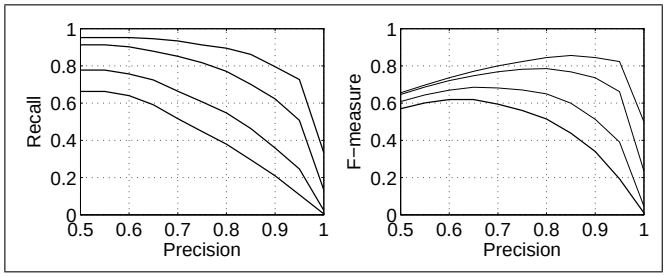

Figure 3. Recall and F-measure as a function of precision for the proposed method with LDA. The four curves from top to down correspond to polyphonies 1, 2, 4, and 6.

Figure 2 shows precision, recall, and F-measure for the proposed enhanced salience function $z(\tau)$ and, for comparison, for the raw salience function $s(\tau)$ used in [4] (here the iterative pitch detection and cancellation was not employed). The results were obtained by fixing a threshold value T, picking all the peaks above the threshold from all the frames, and then calculating the resulting precision π, recall ρ, and F-measure $\varphi = 2\pi\rho/(\pi + \rho)$. As can be seen, the proposed method improves significantly over the baseline method.

Figure 3 shows how the recall and F-measure behave in different polyphonies, varying the number of concurrent sounds from 1 to 6. The number of concurrent sounds in the mixtures was not given, but a single threshold value was again used (common to all polyphonies) and peaks above the threshold were picked from the enhanced salience function.

4 CONCLUSIONS

The proposed method for calculating pitch salience improved significantly over the baseline method in simulations. Furthermore, LDA reduces the feature vector dimensionality to one and does not require more than one Gaussian in the GMM. This means that the proposed method is computationally efficient and can be applied at all points of the raw salience function $s(\tau)$, not only at the positions of the main peaks. This is particularly useful for smooth pitch content visualization and feature extraction purposes.

5 REFERENCES

[1] C. Yeh, *Multiple fundamental frequency estimation of polyphonic recordings*, Ph.D. thesis, University of Paris VI, 2008.

[2] A. de Cheveigné, "Multiple F0 estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. Wang and G. J. Brown, Eds. Wiley–IEEE Press, 2006.

[3] D. P. W. Ellis and G. Poliner, "Classification-based melody transcription," *Machine Learning*, vol. 65, no. 2–3, pp. 439–456, 2006.

[4] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Intl. Conf. on Music Information Retrieval*, Victoria, Canada, 2006.