

Maximum Likelihood Study for Sound Pattern Separation and Recognition

Xin Zhang, Krzysztof Marasek, Zbigniew W. Ras

Computer Science Department, University of North Carolina, Charlotte, N.C., USA

Multimedia Department, Polish-Japanese Institute of IT, Warsaw, Poland

xinzhang@uncc.edu, kmarasek@pjwstk.edu.pl, ras@uncc.edu

Abstract

The increasing needs of content-based automatic indexing for large musical repositories have led to extensive investigation in musical sound pattern recognition. Numerous acoustical sound features have been developed to describe the characteristics of a sound piece. Many of these features have been successfully applied to monophonic sound timbre recognition. However, most of those features failed to describe enough characteristics of polyphonic sounds for the purpose of classification, where sound patterns from different sources are overlapping with each other. Thus, sound separation technique is needed to process polyphonic sounds into monophonic sounds before feature extraction. In this paper, we proposed a novel sound source separation and estimation system to isolate sound sources by maximum likelihood fundamental frequency estimation and pattern matching of a harmonic sequence in our feature database.

1. Introduction

Numerous successful features have been developed to describe the characteristics of monophonic sound pieces. Recently, the Moving Picture Expert Group (MPEG) has published the MPEG7 standard of a set of acoustical features based on latest research in this area. However, most of these features failed to describe enough information to distinguish timbers for polyphonic sounds, where multiple sound sources are active at the same time. Thus, Blind Sound Separation is needed to preprocess polyphonic sounds into monophonic sounds before feature extraction.

Human hearing perception system can focus on a few sound sources in a multi-sounds environment, where different musical instruments are playing at the same time. However, it is a very challenging task for computer to recognize pre-dominant musical sound sources in sound mixtures, which is also called a

Cocktail Party Problem [8]. Next, this paper will contribute a review on Blind Signal Separation and Multi-pitch estimation in the rest of this section, since this work has implications for research in blind harmonic sound separation and pre-dominant fundamental frequency estimation.

1.1. Blind signal separation

Blind Signal Separation is a very general problem in a number of areas besides musical sound timbre recognition: neural computation, finance, brain signal processing, general biomedical signal processing and speech enhancement, etc. Numerous overlapping techniques have been investigated in this area, which can be categorized into, but not limited to the following types: Filtering Techniques ([1] and [4]), Independent Component Analysis (ICA) ([13], and [6]), the Degenerate Un-mixing Estimation Technique (DUET) [14], Factorial Hidden Markov Models (HMM) [16], Singular Value Decomposition (Spectrum Basis Functions in MPEG7 and Harmonic Sources Separation Algorithms ([10] and [18])). Filtering Techniques, ICA and DUET require different sound sources to be stored separately in multiple channels. Most often, HMM works well for sound sources separation, where fundamental frequency range is small and the variation is subtle. However, unfortunately, western orchestral musical instruments can produce a wide range of fundamental frequencies with dynamic variations. Spectral decomposition is used to efficiently decompose the spectrum into several independent subspaces [7] with smaller number of states for HMM. Commonly, Harmonic Sources Separation Algorithms have been used to estimate sound sources by detecting their harmonic peaks, decoding spectrum into several streams and re-synthesizing them separately. This type of methods relies on multi-pitch detection techniques and iterative Sinusoidal Modeling (SM) [10]. For the purpose of interpolating the breaks in the sinusoidal component trajectories, numerous mathematical models have been

explored: linear models [19], and non-linear models such as high degree interpolation polynomials with cubic spine approximation model [10], etc. However, it is very difficult to develop an accurate sinusoidal component model to describe the characteristics of musical sound patterns for all the western orchestral instruments. In this research, we focus on separating harmonic sound signal mixtures in a single channel by isolating and matching the pre-dominant harmonic features with connection to a feature database. In terms of applying harmonic peak information to distinguish timbre, our sound separation method is similar to the SM approach. However, instead of using a model to describe an input signal, we estimate the signal by matching it with the most similar pattern in a harmonic peak feature database. Given an unknown sound mixture, our sound separation system first identifies pre-dominant fundamental frequency among a set of harmonic candidates by a robust maximum likelihood algorithm, and then compares a sequence of its corresponding main harmonic peaks with the ones in our feature database and estimates the unknown sound source by the best match, and then subtracts the matched sound from the unknown sound mixture, and repeats the same steps to the remaining signal.

1.2. Multi-pitch recognition

Pitch detection has been extensively explored by lots of audio signal processing researchers [15] [20] [2] [9]. Pitch detection techniques have been widely used in music transcription and music file annotation. Numerous methods of pitch detection have been developed and explored, which can be categorized by the functional domain into three different types: time, frequency, and time-frequency. This paper focuses on reviewing the most promising type of the fundamental frequency estimation algorithms, which leads to multi-pitch detection: the frequency domain pitch estimation. Since, most famous and well-established algorithms in other domains such as the autocorrelation [2] and the Average Magnitude Difference Function [9] in the time domain, which have been successfully applied in mono-signal processing, fail to detect the fundamental frequencies of sound mixtures in polyphonic sounds.

Many interesting methods have been explored by lots of researchers to detect fundamental frequency in the frequency domain ([5], [12], [3], [2] and [11]). The diagram of a common frequency domain pitch detector is shown in the above figure. One approach is to use a group of hypothetical fundamental frequencies for a comb function [5], where the fundamental frequency is estimated by a hypothetical fundamental frequency that maximized the value of a sum of products A_c of the

comb function and its corresponding power in the STFT spectrum, see formula 1.).

$$C(m, f_h) = \begin{cases} 1: m = kf_h, k \in [1, N] \\ 0: m \neq kf_h \end{cases} \quad (1.)$$

$$A_c(f_h) = \sum_{k=1}^{\frac{N}{2f_h}} X(kf_h) \times C(kf_h, f_h)$$

where kf_h is the k^{th} harmonic of the h^{th} candidate frequency, N is the sampling rate, and X is the power of the spectrum.

Beauchamp et al. extended this algorithm by replacing the comb function with a two-way mismatch function [2].

$$e_1 = \sum_x^{\frac{N}{2f_h}} \min_i |if_h - f_x| \rho(f_x, A_x) \quad (2.)$$

$$e_2 = \sum_x^M \min_i |if_h - f_x| \rho(f_x, A_x)$$

$$E = w_1 e_1 + w_2 e_2$$

where N is the sampling rate, w_1 and w_2 are empirical coefficients. The drawback of this type of algorithms is that the selection of a group of hypothetical fundamental frequencies is critical to their system performance and efficiency.

Another approach is based on the Schroeder's histogram method, which uses the maximum value in the Schroeder's histogram of the integer multiples of each peak frequency to estimate the fundamental frequency [17]. Hess extended this approach by applying a compressed spectrum to the histogram [12]. Edgar et al. improved this algorithm with a maximum likelihood function by taking the distance between the real peak and the integer multiple of a candidate fundamental frequency and the priority of the frequency order into account [3].

$$f_0 = \max \left\{ \sum_{i=1}^k (C \log A_i) \left(C_e \frac{d_i^2 + f_i}{D} \right) \right\} \quad (3.)$$

The above review is not a complete for all the fundamental frequency estimation. It focuses on the pitch detection by the frequency components in the power spectrum. We proposed a robust pre-dominant fundamental frequency algorithm based on the maximum likelihood of the frequency components concept. The following sections begin with an outline of our system, and then describe the details of algorithm in this research.

2. Harmonic signal isolating system

Our system consists of five modules: a quasi-steady state detector, a STFT converter with hamming window, a pre-dominant fundamental frequency

estimator, a sequential pattern matching engine with connection to a feature database, a FFT subtraction device. The quasi-steady state detector computes overall fundamental frequency in each frame by a cross-correlation function, and outputs the beginning and end positions of the quasi-steady state of the input sound. In spectrum from the STFT, the pre-dominant fundamental frequency estimator identifies all the possible harmonic peaks, computes the likelihood value for each candidate peak, elects the frequency with the maximum likelihood value as the fundamental frequency, and passes its normalized correspondence harmonic sequence to the sequential-pattern matching engine, which computes the distance of each pair wise sequence of first N harmonic peaks, then outputs the sound with the minimum distance value for each frame, and finally estimates the sound object by the most frequent sound object among all the frames. The FFT subtraction device subtracts the detected sound source from the spectrum, computes the imaginary and real part of the FFT point by the power and phase information, performs IFFT for each frame, and outputs resultant remaining signals into a new audio data file.

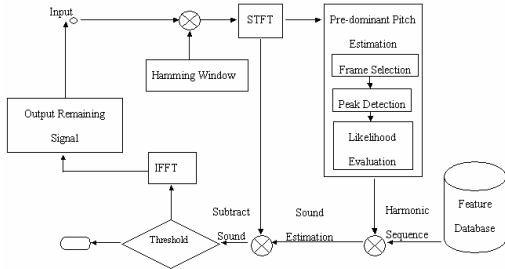


Figure 1. Sound separation system overview.

2.1. Quasi-steady state estimation

This research investigates harmonic sequence information for the purpose of distinguishing the sound timbre, where energy is significantly distributed in harmonic peaks and fundamental frequency variation is relatively subtle. Also, by focusing on the steady frames, it efficiently shrinks down the size of the feature database for the purpose of pattern matching.

$$r(i, k) = \frac{\sum_{j=m}^{m+n-1} s(j)s(j-k)}{\left(\sum_{j=m}^{m+n-1} s(j)^2 * \sum_{j=m}^{m+n-1} s(j-k)^2 \right)^{0.5}} \quad (4.)$$

where s is the audio signal sample data, n represents the frame size, k represents a lag.

The beginning of the quasi-steady state is at the first frame having an overall fundamental frequency in the same frequency bin as its N continuous following neighbor frames, where the total energy in the

spectrum is bigger than a threshold in a case of salience or noise. Each frequency bin corresponds to a music note. The overall fundamental frequency is estimated by pattern recognition with a cross-correlation function [21].

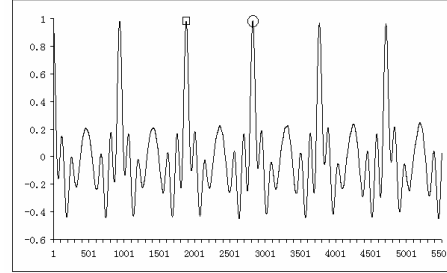


Figure 2. Estimating fundamental frequency by the cross-correlation pattern.

The above figure shows the cross-correlation pattern of a sound played by an electronic bass in a quasi-steady state. The fundamental frequency is computed as the difference between the frequency having the maximum value and the closest frequency having a local maximum value precedent to it where the maximum value is marked by a circle and the precedent local maximum value is marked by a rectangle. An empirical flexible threshold related to the maximum peak of the whole pattern is used to detect a local maximum peak.

Also, points before the first lag where the function value begins to increase are skipped since for low pitch signals, the duration is often relatively insignificant comparing to the whole periodicity and therefore may have the maximum function value.

2.2. Pre-dominant fundamental frequency

In each steady frame, the pre-dominant fundamental frequency is elected among a group of harmonic peaks by a maximum likelihood function. A peak is defined as a point having power value bigger than its immediate neighbor FFT points. Harmonic peaks are estimated by a convolution window of mean amplitude, which is larger than a flexible threshold t .

$$P_i > t, t = C \cdot A_{max}, \chi_i^p > \chi_{i-1}, \chi_i^p > \chi_{i+1} \quad (5.)$$

where C is a empirical coefficient, and A_{max} is the largest amplitude in the spectrum, P_i is the i^{th} candidate peak, χ_i is the power of P_i , χ_{i-1} is the power of the FFT point before P_i , and χ_{i+1} is the power of the FFT point after P_i .

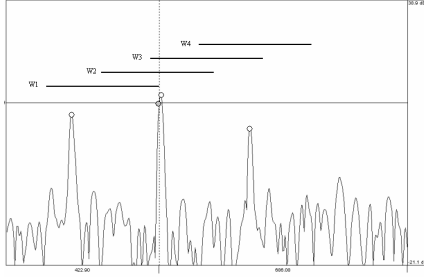


Figure 3. Candidate peaks identification.

The figure above shows that four peak candidates were identified by four continuous windows, one peak per window. The peak marked with gray circle from window1 (displayed as w1 in the figure) is then removed from the candidate list, since it doesn't meet the requirement in the above equation of power: its power is less than one of its immediate neighbor FFT points. Each selected harmonic peak is then treated as a fundamental frequency candidate, where for each candidate, only harmonic peaks in higher ordinal position will be considered as its possible corresponding harmonic peaks. This way, candidate peaks in lower ordinal positions automatically have higher priority to gain accumulative weights. For each candidate peak, the weight W is computed by the following equation.

$$W = \sum_i^{S_r/f_i} 10 \log_{10}(A'_i), \quad A'_i = \max\{A_k\}, k \in [i-c, i+c] \quad (6.)$$

where S_r is the sampling rate, f_i is the frequency of the candidate peak, and c is a range of the possible corresponding harmonic peak. The amplitude of each harmonic peak is normalized by the summation of those of the first N harmonic peaks.

2.3. Sequential pattern matching

After the system detects the pre-dominant fundamental frequency, it queries to the feature database based on this value.

Different music instruments may have very different energy distribution among its harmonic peaks. Some percussive instruments, such as piano and xylophone, have most energy concentrated on a single harmonic peak; some reed instruments, such as flute and horn, have energy more evenly distributed on lots of harmonic peaks; some string instrument, such as violin and cello, have energy concentrated on the first few harmonic peaks. However, little energy distributes on harmonic peaks in higher ordinal position than the tenth harmonic peak.

Due to this fact, our research focuses on the dense energy region of a spectrum by applying an empirical threshold. The distance of a pair of sequences is

measured by a weighted and normalized difference between each peak.

$$D_m(x, y) = \sum_{i=1}^N \frac{x_i |x_i^2 - y_i^2|}{(x_i^2 + y_i^2)X}, \quad X = \sum_{i=1}^N x_i \quad (7.)$$

where x_i is the i^{th} element of the harmonic peak sequence from the m^{th} frame of an unknown musical sound object, y_i is the i^{th} element of the harmonic peak sequence of a sound frame record from the feature database. The best K matching sound frames are chosen where the minimum distance is reached. After repeat the matching procedure for every frame of the unknown musical sound object, $K \times M$ sound frame records are selected, where each group of K records is according to a frame in the unknown sound object, and may belong to different sound records. The sound record having the maximum total number of matched frames is selected as the matching sound object of the unknown sound object, if the total number is above a threshold.

2.4. Sound Subtraction

The harmonic sequence of estimated sound from the database is subtracted from the unknown sound by the real part. The imaginary part is then computed by the phase information of the input unknown sound.

By an IFFT transform and inverse of hamming window, the subtracted spectrum information is projected onto the time domain. Due to the overlapping of the analysis windows, there are duplications in the output. Different duplication-removing and zero-padding methods were compared: one is to accumulate the output of the overlapping area and apply a mean value; others include outputting only one third of the analysis window in the front part, the middle part, and the ending part. Due to the overlapping of two thirds of the analysis window, only one third of the data energy of all the concurrent analysis windows is output to the form a new sound.

2.5. Feature database

Underlying the system is a large feature database, which contains the harmonic peak sequences of sounds originated from the McGill University Master Samples (MUMS) in form of AU format. Every musical sound had multiple records to describe its harmonic information in the frames. A last frame of a sound file, in which the total number of samples is less than the frame size, was truncated, due to the fact that it may not contain enough information to correctly describe the precocity pattern, even though it may be in the steady state. These harmonic peak sequences were

grouped by the corresponding frequency bins of their music notes. Their indices include an audio file name, a frame number, and a peak identification number. The database covers the entire pitch range of all its music instruments.

3. Experiments and results

In this research, sound separation experiments were performed on two different types of sounds: percussive sounds, such as and xylophone, etc., and harmonic sounds, such as guitar, violin, flute and so on. The harmonic sound class contains musical sounds from two different instrument families: string and wind. Each instrument family contained multiple instruments with all different articulations. Totally, there are 97 different music notes in the database, where each musical note was played by a group of different musical instruments. 12 digital sound mixes were made out of 24 randomly selected sounds, which originated from the MUMS. Each pair wise sound came from a different instrument family. During sound mixing, to produce pre-dominant sound source in a sound mixture, one sound signal was reduced to half of its original volume, while the other one was reduced to one eighth of its original volume.

The sampling rate is 44,100Hz, which is a common rate in musical compact CDs. Each sound mix contained two different sound sources, where sample values in one of the sound sources were reduced to one fourth of their original values to make the other sound source pre-dominant. To cover the full range of music notes in our audio database, we used a frame size of 120 milliseconds and a hop size of 40 milliseconds. The feature database contains 3737 sounds having corresponding music notes.

We observed that outputting the middle part of the analysis window achieves the smoothest waveform,

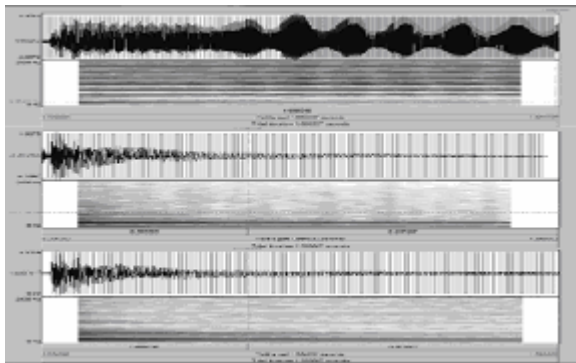


Figure 4. Signal before/after separation.

where a wrapping-up zero-padding method is applied to the FFT and IFFT transform. Finally, a convolution of mean sample value is applied to each ends of an output segment to smooth the waveform. Figure 4 illustrates sound signal waveform before and after separation, where the top image is from a sound mixture of a sound played by a alto-flute in the third octave B and a sound played by cello in the second octave A, the middle one is from the remaining signal after subtracting the B3 also flute signal, and the bottom one is from original MUMS cello sound in the second octave A for comparison. After the convolution window being added, the waveform becomes smoother.

Table1 shows the performance of pitch estimation in our system. The system correctly identified most of the music notes and their corresponding octaves for the polyphonic sounds. Table2 shows the accuracy of pattern matching of the harmonic sequence. Harmonic sounds played by the woodwind instrument and the string instrument generally had higher estimation accuracy in timbre estimation and articulation estimation. Table3 shows the result of the estimation for the remaining in the sound after the sound separation. The performance of Instrument family type estimation was significantly better than that of the individual category estimation. Generally, the articulation estimation had less accuracy than others. Convolution window efficiently removed the jitters in the output waveform.

Table 1. Estimate predominant pitch

	Note	Octave
Perc. & Harm.	66.7%	100%
String	100%	83.3%
Woodwind	100%	75.0%

Table 2. Estimate predominant timbre

	Instr. Family	Instr. Type	Articulation
Perc. & Harm.	100%	83.3%	33.3%
String	100%	100%	83.3%
Woodwind	100%	75.0%	75.0%

Table 3. Estimate pitch of signal residuals

	Note	Octave
Perc. & Harm.	66.7%	66.7%
String	100%	83.3%
Woodwind	100%	100%

Table 4. Estimate timbre of signal residuals

	Instr. Family	Instr. Type	Articulation
Perc. & Harm.	66.7%	66.7%	33.3%
String	100%	100%	83.3%
Woodwind	100%	75.0%	100%

4. Conclusion and future trends

It is possible to estimate the timber of a predominant sound source in a polyphonic sound of different pitches by pattern matching of the harmonic sequential information in a feature database. The weighted dissimilarity measurement of the pattern-matching algorithm can be further improved together with segmentation techniques and convolution masks. Pattern matching of harmonic peak sequence in different states separately provides efficient access to the feature database, and significantly improves accuracy.

5. Acknowledgment

This work is supported by the National Science Foundation under grant IIS-0414815.

6. References

- [1] Balan, R. V., Rosca, J. P., and Rickard, S. T. Robustness of parametric source demixing in echoic environments, in Proc. Int. Conf. on *Independent Component Analysis and Blind Source Separation (ICA)*, 2001, pp. 144-148.
- [2] Beauchamp, J. W., Maher, R.C., and Brown, R. (1993). Detection of Musical Fundamental frequency from Recoded Solo Performances. *94th Audio En. Society Convention*, preprint 3541, Berlin, March 16-19.
- [3] Berdahl, E. and Burred, J.J. (2002) Moderne Methoden der Signal-analyse Abschlußbericht Grund-frequenz-analyse musikalischer Signale, Technische Universität Berlin - FG Kommunikation-swissenschaft, SoSe.
- [4] Brown, G. J., and Cooke, M. P. (1994) Computational auditory scene analysis, *Computer Speech and Language*, vol. 8, pp. 297-336.
- [5] Brown, J.C. (1992). Musical Fundamental Frequency Tracking Using a Pattern Recognition Method. *J. Acoust. Society Am.* 92(3), 1394-1402.
- [6] Cardoso, J.F. (1998) Blind source separation: statistical principles, *Proceedings of the IEEE*, vol. 9, no. 10, pp.2009-2025.
- [7] Casey, M. A., and Westner, A. (2000) Separation of mixed audio sources by independent subspace analysis, in Proc. *International Computer Music Conference (ICMC)*, pp. 154-161.
- [8] Cherry, E. Collin. (1953) Some Experiments on the Recognition of Speech, with One and with Two Ears. *Journal of the Acoustical Society of America*, 24, pp. 975-979.
- [9] Cook, P.R., Morill, D., and Smith, J. O. (1998). An Automatic Pitch Detection and MIDI Control System For Brass Instruments. *J. Acoustics Society Am*, 92(4 pt. 2), 2429-2430.
- [10] Dziubinski, M., Dalka, P., Kostek, B (2005) Estimation of Musical Sound Separation Algorithm Effectiveness Employing Neural Networks, *Journal of Intelligent Information Systems*, 24(2/3), 133-158.
- [11] Goto, M. (2000). A Robust Predominant-F0 Estimation Method for Real-Time Detection of Melody and Bass Line in CD Recordings. In Proc. *IDDD International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, pp.11-757-760.
- [12] Hess, W. (1983). Fundamental frequency Determination of Speech Signals: Algorithms and Devices. Springer Berlin: Verlag, Tokyo: Heidelberg, New York.
- [13] Hyvarinen, A., Karhunen, J. and Oja, E. (2001). Independent Component Analysis. John Wiley & Sons, 2001.
- [14] Jourjine, A. N., Rickard, S. T. and Yilmaz, O. (2000). Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures, in Proc. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. V-2985-2988.
- [15] Klapuri, A. (1999). Wide-band Pitch Estimation for Natural Sound Sources with In-harmonicities. *106th Audio Engineering Society Convention*, Preprint 4906, Munich, May 8-11.
- [16] Ozerov, A., Philippe, P., Gribonval, R. and Bimbot, F. "One microphone singing voice separation using source adapted models", in Proc. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 90-93.
- [17] Schroeder, M.R. (1968). Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurement. *J. Acoust. Society Am.*, 43, 829-834.
- [18] Smith, J.O. and Serra, X. (1987). PARSHL: An Analysis/Synthesis Program for Non Harmonic Sounds Based on a Sinusoidal Representation. In Proc. *Int. Computer Music Conf.* (pp.290-297), Urbana-Champaign, Illinois.
- [19] Virtanen, T. and Klapuri, A. (2000) Separation of Harmonic Sound Sources Using Sinusoidal Modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey.
- [20] Walmsley, P.J., Godsill, S.J., and Rayner, P.J.W. (1999). Polyphonic Pitch Tracking Using Joint Bayesian Estimation of Multiple Frame Parameters. In *IEEE Workshop on Applications of signal Processing to Audio and Acoustics*, 17th-20th October: New Paltz (NY).
- [21] Zhang, X. and Ras, Z.W. (2006A). Differentiated Harmonic Feature Analysis on Music Information Retrieval for Instrument Recognition, proceeding of *IEEE International Conference on Granular Computing*, May 10-12, Atlanta, Georgia, 578-581.