MULTIPITCH ESTIMATION AND SOUND SEPARATION BY THE SPECTRAL SMOOTHNESS PRINCIPLE

Anssi P. Klapuri

Tampere University of Technology, P.O.Box 553, FIN-33101 Tampere, Finland
klap@cs.tut.fi

ABSTRACT

A processing principle is proposed for finding the pitches and separating the spectra of concurrent musical sounds. The principle, spectral smoothness, is used in the human auditory system which separates sounds partly by assuming that the spectral envelopes of real sounds are continuous. Both theoretical and experimental evidence is presented for the vital importance of spectral smoothness in resolving sound mixtures. Three algorithms of varying complexity are described which successfully implement the new principle. In validation experiments, random pitch and sound source combinations were analyzed in a single time frame. Number of simultaneous sounds ranged from one to six, database comprising sung vowels and 26 musical instruments. Usage of a specific yet straightforward smoothing operation corrected approximately half of the pitch errors that occurred in a system which was otherwise identical but did not use the smoothness principle. In random four-voice mixtures, pitch error rate reduced from 18% to 8.1%.

1. INTRODUCTION

Pitch perception plays an important part in human hearing and understanding of sounds. In an acoustic environment, human listeners are able to perceive the pitches of several simultaneous sounds and make efficient use of the pitch to "hear out" a sound in a mixture [1]. Computational modeling of this function, multipitch estimation, has been relatively little explored in comparison to the availability of algorithms for single pitch estimation in monophonic speech signals [2].

Until these days, computational multipitch estimation (MPE) has fallen clearly behind humans in accuracy and flexibility. First attempts were made in the field of automatic transcription of music, but were severy limited in regard to the number of simultaneous sounds, pitch range, or variety of sound sources involved [3]. In recent years, further progress has taken place. Martin proposed a system that utilized musical knowledge in transcribing four voice piano compositions [4]. Kashino *et al*. describe a model which was able to handle several different instruments [5]. Goto's system was designed to extract melody and bass lines from real-world musical recordings [6]. Psychoacoustic knowledge has been succesfully utilized e.g. in the models of Brown and Cooke [7], Godsmark *et al*. [8], and de Cheveigne and Kawahara [9]. Also purely mathematical approaches have been proposed [10].

Multipitch estimation and auditory scene analysis are intimately linked. If the pitch of a sound can be determined without getting confused by other co-occurring sounds, the pitch information can be used to organize simultaneous spectral components to their sources of production. Or, vice versa, if the spectral components of a source can be separated from the mixture, MPE reduces to single pitch estimation. This is why most recent MPE systems explicitly refer to and make use of the human auditory scene analysis principles. In human hearing, the perceptual organization of spectral components has been found to depend on certain acoustic cues. Two components may be associated to a same source by their closeness in time or frequency, harmonic concordance, synchronous changes in the frequency or amplitude of the components, or spatial proximity in the case of multisensor input [1].

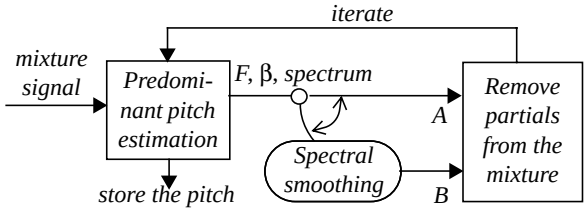

Fig. 1. Experimental framework: system which can be switched between two modes. (*A*) Straightforward iterative approach. (*B*) Spectral-smoothness based model.

The purpose of this paper is to propose a new efficient mechanism in computational MPE and auditory organization. *Spectral smoothness* refers to the expectation that the spectral envelopes of real sound sources tend to be continuous. Bregman points out this principle in human hearing by mentioning that spectral smoothness promotes integration of frequency partials to a same source and a single higher intensity partial is more likely to be perceived as an independent sound [1, p.232]. However, smoothness has not traditionally been included among the auditory organization cues. This paper presents evidence for the importance of spectral smoothness both in human and computational MPE. Also, three different algorithms are described that implement this principle.

Validation experiments were performed using an experimental model, where the spectral smoothness was either utilized in different ways, or was completely ignored. Acoustic database comprised sung vowels and the whole pitch range of 26 musical instruments. MPE was performed in a single time frame for random pitch and sound source combinations, number of simultaneous sounds ranging from one to six. Including the spectral smoothness principle in calculations made significant improvement in simulations. For example, the pitch error rate in random four-voice mixtures dropped from 18 % to 8.1 %, and in musical four-voice mixtures from 25 % to 12 %. As a result, MPE could be performed quite accurately at a wide pitch range and without *a priori* knowledge of the sound sources involved.

2. EXPERIMENTAL FRAMEWORK

Figure 1 shows the overview of the system which acts as an experimental framework in this paper. The system can be

switched between two modes. The straightforward iterative MPE model, denoted by branch *A*, has been described earlier in [11]. It consists of two main parts that are applied in an iterative succession. The first part, predominant pitch estimation, finds the pitch of the most prominent sound in the interference of other harmonic and noisy sounds. As an output, it gives the fundamental frequency *F*, inharmonicity factor β, and the precise frequencies and amplitudes of the harmonic partials of the sound. In the second part, the spectrum of the detected sound is linearly subtracted from the mixture. These are then repeated for the residual signal.

A spectral-smoothness based model is obtained by locating an additional module between the estimation and subtraction stages. This is denoted by branch *B* in Fig. 1. The aim of the spectral smoothing algorithm is to use the pitch information to produce a more appropriate estimate for the spectrum of a separated sound before it is subtracted from the mixture. The need for such a module is strongly motivated by two observations. The predominant pitch estimation algorithm is capable of finding one of the correct pitches with 99 % certainty even in six-voice polyphonies [11]. However, the probability of error increases rapidly in the course of iteration. This indicates that the initial estimate of a sound spectrum as given by predominant pitch algorithm is not accurate enough to remove it correctly from the mixture.

3. DIAGNOSIS OF THE STRAIGHTFORWARD ITERATIVE SYSTEM

Simulations were run to analyze the behaviour of the straightforward iterative estimation and separation approach, i.e., the branch *A* in Figure 1. Random mixtures of *N* sounds were generated by first allotting an instrument and then a random note from its whole playing range, however, restricting the pitch over five octaves between 65 Hz and 2100 Hz. The desired number of sounds was allotted, and them mixed with equal mean square levels. The iterative process was then evoked and requested to extract *N* pitches from the acoustic mixture signal. As a general impression, the presented iterative approach works rather reliably.

However, an important observation is immediately made when the distribution of the remaining errors is analyzed. Figure 2 shows the errors as a function of the musical intervals that occur in the erroneously transcribed sound mixtures. It appears that the error rate is strongly correlated with certain pitch relations. More exactly, the straightforward estimation and subtraction approach is likely to fail in cases where the fundamental frequencies of simultaneous sounds are in simple rational number relations, also called *harmonic* relations. These are indicated over the corresponding bars in Fig. 2.

3.1 Coinciding sinusoidal partials

It turned out that coinciding frequency partials from different sounds make the algorithm fail. If sounds are in a harmonic relation to each other, a lot of partials coincide, i.e., share the same frequency. When the firstly detected sound is removed, the coinciding harmonics of remaining sounds are also removed in the subtraction procedure. In some cases, and particularly after several iterations, a remaining sound gets too corrupted to be correctly analyzed in the coming iterations.

When two sinusoidal partials with amplitudes a_1 and a_2 and

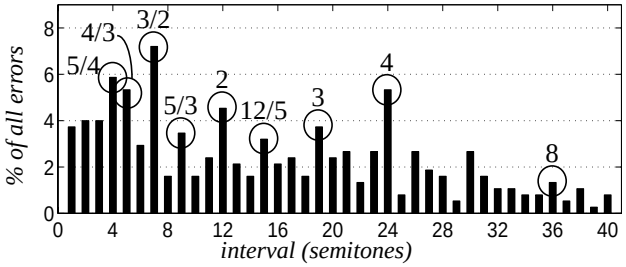

Fig. 2. Distribution of the pitch estimation errors as a function of the musical intervals that occur in the erroneously transcribed sound mixtures.

phase difference θ_Δ coincide in frequency, the amplitude of the resulting sinusoid can be calculated as

$$a_s = \left| a_1 + a_2 e^{i\theta_\Delta} \right| . \quad (1)$$

If the two amplitudes are roughly equivalent, the partials may either amplify or cancel each other, depending on θ_Δ. However, if one of the amplitudes is significantly larger than the other, as is usually the case, a_s approaches the maximum of the two.

3.2 Fundamental frequency relations

The condition that a harmonic partial *h* of a sound *S* coincides a harmonic *j* of another sound *R* can be written as $hF_S = jF_R$, where F_S and F_R are the fundamental frequencies, and the two sides represent the frequencies of the partials. When the common factors of integers *h* and *j* are reduced, this yields

$$F_R = \frac{m}{n} F_S , \quad (2)$$

where $(m, n) \geq 1$ are integer numbers. This implies that partials of two sounds can coincide only if the fundamental frequencies of the two sounds are in rational number relations. Furthermore, when the fundamental frequencies of two sounds are in the above relation, then every m^{th} harmonic *mk* of the sound *S* coincides every n^{th} harmonic *nk* of the sound *R* at their common frequency bands, where integer $k \geq 1$. This is evident since hF_S equals jF_R for each pair *h=mk* and *j=nk*, when Eq. (2) holds.

An important principle governing music is paying attention to the pitch relations, intervals, of simultaneously played notes. Simple harmonic relations satistying Eq. (2) are favoured over dissonant ones. Although western music arranges notes to a quantized logarithmic scale, it can surprisingly well produce the different harmonic intervals that can be derived by substituting small integers to Eq. (2) [3]. Because harmonic relations are so common in music, these “worst cases” must be handled well in general. Also, this explains why MPE is particularly difficult in music.

4. SOLUTION AND ITS ARGUMENTATION

The difficulties caused by harmonic pitch relations can be classified into two categories. First, the partials of an other sound may be erroneously removed along with the one that is being actually separated. This causes undetections. Second, two or more fundamental frequencies in certain relations may make a non-existent “ghost” sound appear, for example the root pitch of a chord. This causes insertion errors, i.e., extraneous pitch detections.

There is a solution to these problems that is both intuitive,

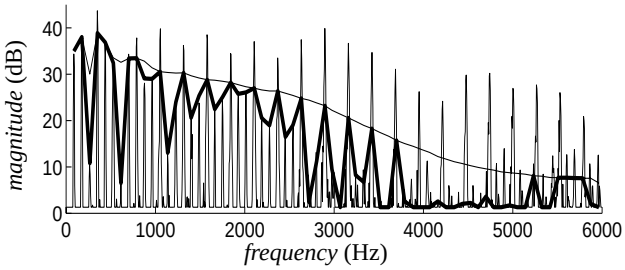

Fig. 3. Illustration of the spectral smoothness principle. Logarithmic magnitude spectrum containing two sounds, lower of which has been detected first. The spectrum has been high-pass liftered to remove spectral envelope.

efficient, and psychoacoustically valid: the spectra of the detected sounds must be smoothed before subtracting them from the mixture. Consider the logarithmic magnitude spectrum of a two-sound mixture in Fig. 3. The harmonic partials of the higher-pitched sound coincide every third harmonic of the lower-pitched sound, which has been detected first. As predicted by Eq. (1), the coinciding partials of the detected sound tend to have higher magnitudes than the other ones. However, when the sound spectrum is smoothed (thin slowly decreasing horizontal curve in Fig 3), these partials rise above the smooth spectrum, and thus remain in the residual after subtraction. In this way, the other sound is not removed with the detected one. When properly applied, the same mechanism can be used to treat ghost sounds, too.

4.1 Psychoacoustic knowledge applied

The design of the smoothing operation is not as simple as it seems to be at the first glance. As a matter of fact, simply smoothing the amplitude envelope (thin horizontal curve in Fig 3) before subtraction from the mixture does *not* work in the sense that it would reduce the pitch error rate in simulations.

Spectral smoothing in the human auditory system does not take the form of lowpass filtering. Instead, a nonlinear mechanism cuts off single higher amplitude partials. In following, a brief description of the human auditory processing is made in order to reveal the exact mechanism of an appropriate smoothing process.

Meddis and Hewitt have proposed a computer model of human auditory periphery which aims at reproducing a widest range of phenomena in human pitch perception [12]. The algorithm consists of four main steps. First, the input signal is passed through a bank of bandpass filters. At each band, the signal is halfwave rectified and lowpass filtered to extract the amplitude envelope of the bandpassed signal. Periodicity in the resulting signal is detected by calculating autocorrelation function estimates within channels. In the final phase, the estimates are linearly summed across channels to get a summary autocorrelation function, the maximum value of which points out the global pitch.

Amplitude envelope calculation within channels performs implicit spectral smoothing. When a harmonic sound is considered, each two neighbouring harmonic partials cause amplitude *beating*, i.e., alternatingly amplify and cancel each other at the fundamental frequency rate. However, the magnitude of the beating caused by each two sinusoidal partials is determined by the smaller of their amplitudes. When the spectrum of a harmonic sound is considered, this “minimum amplitude” property filters out single higher amplitude harmonic partials.

4.2 Three smoothing algorithms

A computer implementation of the implicit smoothing in the human auditory system can be isolated to a separate module. The algorithm simply goes through the harmonic partials of a sound and replaces the amplitude a_h of partial h with the minimum of the amplitudes of the partial and its neighbour

$$a_h \leftarrow min(a_h, a_{h+1}) . \quad (3)$$

Interestingly, performing this simple operation in the spectral smoothing module of Fig. 1 corrects about 30 % of the errors of the straightforward iterative model. For example, the error rate in random four-voice mixtures reduces from 18 % to 12 %.

A still more efficient algorithm can be designed by focusing on the role of the smoothing algorithm. It is: to cut off single clearly higher amplitude partials. Equation (3) surely does that, but bases the estimate on two values only. The robustness of the method can be improved by imitating the calculations of the human auditory system at bandlimited frequency channels.

The second algorithm first calculates moving average over the amplitudes of the harmonic partials. An octave wide Hamming window is centered at each harmonic partial h, and a weighted mean m_h of the amplitudes of the partials in the window is calculated. This is the smooth spectrum illustrated by a thin horizontal curve in Fig. 3. The original amplitude value a_h is then replaced with the minimum of the original and the averaged amplitude

$$a_h \leftarrow min(a_h, m_h) . \quad (4)$$

These values are illustrated by a thick horizontal curve in Fig. 3. This straightforward algorithm is already almost as good as could be designed. For example, for random four-voice mixtures, the average pitch error rate dropped from 18 % to 8.9 %.

A final slight improvement to the method can be made by utilizing the statistical dependency of every m^{th} harmonic partials, as explained in Sec. 3.2. The third algorithm applies a multistage filter which consists of the following steps. First, the numbers {..., h–1, h, h+1, h+2...} of the harmonic partials around harmonic h are collected from an octave wide window. Next, the surrounding partials are classified into groups, where all the harmonics that share a common divisor are put to a same group. Third, estimates for harmonic h are calculated inside groups in the same manner as in the second algorithm. In the last step, the estimates of different groups are averaged, weighting each group according to its mean distance from harmonic h.

The other problem category, that of ghost sounds, was solved by noticing that that the likelihood of a predominant pitch should be re-estimated after the new smooth spectrum is calculated. An example case clarifies why an erroneous sound may arise as a joint effect of the others and how the problem can be solved. If two harmonic sounds are played with fundamental frequencies 2F and 3F, the spectra of these sounds match every second and every third harmonics of a non-existent sound with fundamental frequency F, which is erroneously credited for all the observed partials, and thus appears as a ghost sound. However, if the harmonic amplitudes of the ghost sound are smoothed and its likelihood is re-estimated, the irregularity of the spectrum decreases the level of the smooth spectrum, and the likelihood remains low.

Table 1: Pitch error rates using different smoothing algorithms.

Applied smoothing algorithm	Random mixtures, four voices	Musical mixtures, four voices
—None—	18 %	25 %
SMOOTH	18 %	24 %
MIN (1st)	12 %	17 %
SMOOTH+MIN (2nd)	8.9 %	13 %
STAT+MIN (3rd)	8.1 %	12 %

5. SIMULATIONS RESULTS

A lot of simulations was run to verify the importance of the proposed spectral smoothness principle and to compare the described three algorithms. Table 1 gives the pitch error rates using different spectral smoothing algorithms. Algorithms are listed top–down in the order they were introduced in this paper. The first row gives the results using the straightforward iterative estimation and separation approach, with no smoothing. Label SMOOTH refers to simple smoothing of the amplitude envelope, which is of no help, as mentioned in Sec. 4.1. MIN refers to the minimum-among-neighbours algorithm implemented by Eq. (3). SMOOTH+MIN is the second algorithm, given by Eq. (4). STAT+MIN is the third algorithm utilizing statistical dependencies between the partials.

Random mixtures were generated in the way described in Sec. 3. In musical mixtures, different pitch relations were favoured according to a statistical profile discovered by Krumhansl in classical western music [13, p.68]. In all simulations, pitch estimation took place in a single 190 ms time frame 100 ms after the onsets of the sounds. A correct pitch was defined to deviate less than half a semitone (±3 %) from the correct value.

As the most important observation, spectral smoothing makes remarkable improvement to MPE accuracy. The third algorithm is slightly but consistently the best, but also by far the most complicated among the three. The second algorithm, while being very simple to implement, already achieves almost same performance.

Figure 4 shows multipitch estimation results in different polyphonies using the STAT+MIN algorithm. The bars represent the overall error rates as a function of the polyphony, where e.g. error rate for random four-voice polyphonies is 8.1 % on average. The different shades of grey in each bar indicate the error cumulation in the iteration, errors occurred in the first iteration at the bottom.

The system works reliably and exhibits graceful degradation in increasing polyphony, with no abrupt breakdown at any point. As predicted by the analysis in Sec. 3.2, musical mixtures were generally more difficult to resolve. However, the difference is not very big, indicating that the spectral smoothing works well.

6. CONCLUSIONS

Spectral smoothness principle was proposed as an efficient new mechanism in MPE and sound separation. Introduction of this idea corrected approximately half of the errors occurring in an otherwise identical system which did not use the smoothness principle. As a result, MPE could be performed quite accurately at a wide pitch range and without *a priori* knowledge of the sound sources involved. The underlying assumption that the spectral

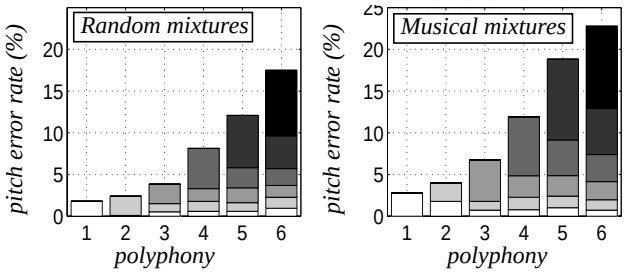

Fig. 4. Pitch error rates for multipitch estimation in different polyphonies. Bars represent the overall error rates, and the different shades of gray the error cumulation in iteration.

envelopes of natural sounds are rather continuous seems to hold, since the smoothing operation can be done without noticeable loss of information from the MPE viewpoint.

7. REFERENCES

[1] Bregman, A. S.(1990). "Auditory Scene Analysis," MIT Press.

[2] Rabiner, L. R., Cheng, M. J., Rosenberg, A. E., and McGonegal C. A. (1976). "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. Acoust., Speech, and Signal Processing, Vol.ASSP-24, No.5, 399–418.

[3] Klapuri, A. P. (1998). "Automatic Transcription of Music," MSc thesis, Tampere University of Technology, 1998.

[4] Martin, K. D. (1996). "Automatic Transcription of Simple Polyphonic Music: Robust Front End Processing", Massachusetts Institute of Technology Media Laboratory Perceptual Computing Section Technical Report No. 399.

[5] Kashino, K., Nakadai, K., Kinoshita, T., and Tanaka, H. (1995). "Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitative Information Integration Mechanism," Proc. International Joint Conf. on Artificial Intelligence, Montréal.

[6] Goto, M. (2000). "A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings," Proc. IEEE International Conf. on Acoust., Speech, and Signal Processing, Istanbul, Turkey.

[7] Brown, G. J., and Cooke, M. P. (1994). "Perceptual grouping of musical sounds: A computational model," J. of New Music Research 23, 107–132.

[8] Godsmark, D., and Brown, G. J. (1999). "A blackboard architecture for computational auditory scene analysis," Speech Communication 27, 351–366.

[9] de Cheveigné, A., and Kawahara, H. (1999). "Multiple period estimation and pitch perception model," Speech Communication 27, 175–185.

[10] Sethares, W. A., and Staley, T. W. (1999). "Periodicity Transforms," IEEE Trans. Signal Processing, Vol. 47, No. 11.

[11] Klapuri, A. P., Virtanen T. O., and Holm, J.–M. (2000). "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals". In Proc. COST-G6 Conference on Digital Audio Effects, Verona, Italy.

[12] Meddis, R., and Hewitt, M. J. (1991). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J. Acoust. Soc. Am. 89 (6).

[13] Krumhansl, C. L. (1990). "Cognitive Foundations of Musical Pitch," Oxford University Press, New York.