CONSTRAINED EM ESTIMATES FOR HARMONIC SOURCE SEPARATION

Pamornpol Jinachitra

Center for Computer Research in Music and Acoustics, Stanford University
Stanford, CA94305, USA
pj97@stanford.edu

ABSTRACT

A constrained iterative method for harmonic source parameter estimation is proposed based on an EM algorithm with an intent for harmonic source separation. The problem of coinciding partials and interference among them in general is mitigated by the constraints on the "weak" partials on the stronger ones of the same harmonic source. A useful scheme to determine the weakness of a partial is proposed. The constrained iteration is shown to give more accurate estimates of the sinusoidal parameters which results in good source separation even in most cases of highly overlapping spectra.

1. INTRODUCTION

Sound source separation based on sinusoid modeling is useful in a recovery of vocal or musical instruments from a single channel record. It relies on accurate estimations and tracking of the parameters in the sinusoidal model, namely, the frequencies, the amplitudes and the phases [1]. A lot of work has been done on such parameter estimation in the case of a single partial. However, with more than one partials, the estimation is not as straightforward due to the interference among the components. Another difficulty is in estimating frequency parameters which is non-linear with respect to the observed signal. An iterative parameter estimation is proposed in [2] by iterating through updated estimates of amplitude-phase and the frequencies in turn. A complex linearization of the Fourier transform of a windowed signal is employed to circumvent the problem of non-linearity in the frequency. Apart from having to deal with an inverse of a complex number matrix which is sometimes ill-conditioned when some partials are close by, the processing window also has to be restricted to a non-sidelobe one. In this paper, an alternative iterative method is then proposed. It is based on an EM algorithm developed in [3] for general parameter estimations of superimposed signals, extended to an estimation of amplitudes, frequencies and phases of a single mixture of multiple harmonic sources. The algorithm attempts to find the maximum-likelihood estimates of those parameters. Its attractiveness lies in the ability to decouple the problem into components which then allows for separate optimization on each set of partial's parameters. However, it often gets confused when some of the partials are coinciding or become close by in frequency and gives poor estimates as a result. Fortunately, the harmonic structure of each source allows us to pool information among them [8]. Constraining the "weak" partials on the stronger ones can then give more accurate results. To decide which partials are weak and hence not so trustworthy, a measure of its interference by other partials can be used. The accuracy of estimations in various cases is reported and used in source separation.

2. EM ITERATIVE ESTIMATION

2.1. Signal model

The observed signal is modelled over a processing frame $t = 0, 1, ..., T-1$ as

$$y(t) = \sum_{k=1}^{K} \sum_{j=0}^{H(k)-1} s_{k,j}(t; \theta_{k,j}) + v(t) \qquad (1)$$

$$s_{k,j} = a_{k,j} \cos(2\pi f_{k,j} t + \phi_{k,j}) \qquad (2)$$

where $\theta_{k,j} = [a_{k,j}, f_{k,j}, \phi_{k,j}]$, K is the number of sources, $H(k)$ is the number of harmonic partials belonging to source k, $a_{k,j}$, $f_{k,j}$ and $\phi_{k,j}$ are the amplitudes, the frequencies and the phases associated with them. $v(t)$ is assumed to be real additive white Gaussian noise. The sources are assumed to be harmonic so that each harmonic frequency of each source is approximately an integer multiple of the source fundamental frequency f_0. The indexing of j is set to reflect the convention for harmonics. It is assumed that the sinusoidal parameters are stationary over the frame. This is acceptable for signal with slowly varying parameters and/or the use of short processing frame.

2.2. EM algorithm on superimposed signals

EM algorithm is widely used to estimate parameters from incomplete data [4]. With an apprporiate choice of the complete data, the parameters can be estimated by maximizing the marginalized expectation of the likelihood over the missing components. The current estimates are then used to find the conditional expectation and the process is reiterated. In the problem of our interest, the incomplete data is the observation $y(t)$ whereas the complete data can be chosen as $x_k(t)$ where $y(t) = \sum_{k=1}^{K} \sum_{j=0}^{H(k)-1} x_{k,j}(t)$, $x_{k,j}(t) = s_{k,j}(t) + v_{k,j}(t)$. Also, $v(t) = \sum_{k=1}^{K} \sum_{j=0}^{H(k)-1} v_{k,j}(t)$ is an arbitrary decomposition of the noise. For convenience, all the noise compoenents are assumed to be statistically independent, zero-mean Gaussian with variance $\sigma_{k,j}^2$ associated with each of them where $\sigma_{k,j}^2 = \beta_{k,j}\sigma^2$ is the fraction of actual noise power assigned to the component.

With some modification from [3], the EM iterative steps become

At i^{th} iteration,
E step : for $k = 1, 2, ..., K$ and for $j = 0, 1, ..., H(k) - 1$, compute

$$\hat{x}_{k,j}(t) = s_{k,j}(t;\theta^{(i)}_{k,j}) + \beta_{k,j}\left[y(t) - \sum_{l,m} s_{l,m}(t;\theta^{(i)}_{l,m})\right] \quad (3)$$

M step : for $k = 1, 2, ..., K$ and for $j = 0, 1, ..., H(k) - 1$,

$$\theta^{(i+1)}_{k,j} \leftarrow \min_{\theta_{k,j}} \sum_{t=0}^{T-1} ||\hat{x}^{(i)}_{k,j}(t) - \sum_{j,k} s_{k,j}(t;\theta)||^2 \quad (4)$$

The decoupling of individual components results from the statistical independence of the decomposed noise components. The likelihood maximization step corresponds to a least squares problem when the noise is Gaussian, to be carried out on each component independently. This can reduce the computational complexity greatly especially when matrix inverse or parameter searching will be involved. The EM algorithm is also gauranteed to converge to a local maximum, though, as in any iterative method, good initialization is needed to ensure the global maximum. Despite the much reduced dimension of the problem space, solving for $\theta_{k,j}$ is still not trivial. However, the theory of conditional EM algorithm(ECM) [5] allows the M-step to be done in many small steps, conditioning on other parameters being fixed while retaining the convergence property of the original EM. The amplitude-phase and the frequency of the partial are hence estimated in separate steps.

Note that $\beta_{k,j}$ is the fraction of noise power assigned to the component set arbitrarily subject to $\sum_{k,j} \beta_{k,j} = 1$. It is possible that we set them to reflect the extent of noise in each component to assist in adaptation. Unless there is a scheme to assign them appropriately, they are set to be equal for fairness.

2.3. Amplitude-phase estimation

Dropping all subscripts on considering a single partial, let

$$\mathbf{A}\vartheta = \mathbf{x} \quad (5)$$

where $\mathbf{x}$ is a frame of estimated partial of length T, $\mathbf{A} = [\mathbf{c}, \mathbf{s}]$ and $\mathbf{c}$ and $\mathbf{s}$ are columns of cosine and sine values for $t = 0, ..., T-1$ of the current frequency estimate respectively. The parameter vector $\vartheta = [a\cos\phi, a\sin\phi]^T$. We can solve for amplitude and phase by simple linear least squares

$$\vartheta = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}\mathbf{x} \quad (6)$$

and then solve for a and ϕ from

$$a = \sqrt{\vartheta_1^2 + \vartheta_2^2}, \qquad \phi = \arctan(\vartheta_2/\vartheta_1) \quad (7)$$

Assuming the adaptation should be smooth, phase unwrapping is used to make sure that the *arctan* function gives the wanted value.

2.4. Frequency estimation

Because of non-linearity of the frequecy with respect to the signal, a close-form solution is not available. A gradient descent is employed as a way to get close to the minimizing value of the least squares. The frequency update at the i^th iteration is

$$\Delta f = \sum_{t=0}^{T-1}\left(x(t) - a\cos\left(\frac{2\pi\hat{f}^{(i)}t}{f_s} + \phi\right)\right) a\sin\left(\frac{2\pi\hat{f}^{(i)}t}{f_s} + \phi\right)\frac{2\pi t}{f_s} \quad (8)$$

The two steps minimization gives a monotonic decrease in sum squared errors and hence a monotonic increase in the likelihood. Though the iteration on the frequency estimate may not give an exact minimizer, depending on the learning rate and number of iterations, the generalized EM theory states that convergence is still gauranteed as long as the likelihood is made monotonically increasing. The algorithm is therefore very robust with respect to convergence. It can be iterated until some specific convergence criterion is met.

3. HARMONICALLY CONSTRAINED EM

Despite an independent treatment on each partial, it has been found that the algorithm given above does not do well when some partials have the same frequency or are very close by. It tends to spread the amplitude among the partials incorrectly which might have been caused by the use of equal β. The frequency estimates of the partials involved are also poor which can in turn bring down the performance on other partials. This situation occurs frequently in Western music where integer ratios of pitch intervals are common or in a rich polyphonic spectrum. In general, solving for nearly coinciding partials is difficult but with a harmonic structure available, reasonably good recovery is possible. The "weak" partials obscured by other neighbouring partials can be harmonically constrained in relation to other "strong" ones of the same source during each iteration. If there are enough strong partials available, the constraints will usually give better overall estimates than letting the algorithm search for them freely. How to define "strong" and "weak" is considered in the next section.

3.1. Credit assignments

There are many ways to assign credibility or trustworthiness to a partial. One useful scheme is to look at the position of the partial in the spectrum and its amplitude. If it is close to a larger partial, its trustworthiness is low. On the other hand, if it is relatively larger and far from others, its estimate from the algorithm should be trustworthy and hence can be used as a reference for other weaker ones of the same source. The score of each partial, motivated by spectral interference, can be calculated as

$$c_{k,j} = \frac{1}{(H-1)} \sum_{l,m \neq k,j} [1 - \bar{a}_{l,m} g(|\hat{f}_{l,m} - \hat{f}_{k,j}|)] \quad (9)$$

where H is the total number of partials involved and $\bar{a}_{l,m} = \hat{a}_{l,m} / \sum_{\forall k,j} \hat{a}_{k,j}$ is the normalized amplitude. The function $g()$ is appropriate spectral envelope approximation of the processing window transform. For example, for a rectangular window, we may use,

$$g(f) = \begin{cases} 0.5 + 0.5\cos(2\pi \frac{fT}{f_s\zeta}), & f < f' \\ \alpha/f, & f > f' \end{cases} \quad (10)$$

where ζ is a suitable scaling. $\zeta = 3$ gives close approximation to the *sinc* envelope in the mainlobe while α ensures a continuity from the main lobe part to the sidelobe $1/f$ roll-off envelope approximation. $f' = f_s/T$ is an appropriate boundary where $1/f$ starts to dominate. A similar expression can be obtain for other window transform by suitable scaling but the closeness may not be crucial as long as the general trend of interference is well-represented. The envelope weighted by the relative amplitude reflects how much the peak of other partials will affect the peak of the partial of interest in the spectrum while ignoring the side-lobe oscillation. Consequently, it reflects the trustworthiness of the partial from estimation, especially when the initial estimation is likely to involve a peak picking process. Perfectly overlapping of peaks causes bad credit score but the larger one is still allowed to have a relatively good score. This is desirable since we do not want to throw away too much information. Similar weighting scheme e.g. exponential weighting should work as well but weighting in dB scale magnitude should not be appropriate since very small ripple will be over-emphasized. Besides, the interference is additive in linear scale, not multiplicative.

The decision rule for weak partials may vary. A threshold maybe set between zero and one or in relation to the maximum of the set. Often, it is found that working with "distrust" defined as $d_{k,j} = 1 - c_{k,j}$ is more amenable to analysis. The decision rule that is found to be effective is then to decide that a partial is weak when its distrust score, d, is more than twice of the lowest. The number of partials of the source can also be taken into account. Rejecting too many partials as weak can reduce its robustness while keeping too many not so strong can also bring the performance down. Clearly, there must be at least one decidedly strong partial per source to be successful.

3.2. Constrained estimation of weak partials

Now that it has been decided which partials are strong enough to take part in the iteration, during each EM iteration, the weak partial j of source k will be updated according to.

$$\hat{f}_j = \sum_{m \in S_k} \frac{j}{m} w_j(\hat{f}_m)\hat{f}_m \qquad (11)$$

where S_k is the set of strong partials of source k. The weighting $w_j(\hat{f}_m)$ is a function of $\hat{f}_m$. If the frequency estimates are uncorrelated, the minimal variance solution for the weighting would be

$$w_j(\hat{f}_m) = \frac{1/\sigma^2_{\hat{f}_m}}{\sum_{m \in S_k} 1/\sigma^2_{\hat{f}_m}} \qquad (12)$$

where $\sigma^2_{\hat{f}_m}$ is the variance of the estimator $\hat{f}_m$. This will give $\sigma^2_{\hat{f}_j} < j^2 \min(\sigma^2_{\hat{f}_m}/m^2)$. It also indicates that good high frequncy harmonic estimates will push down the bound, ignoring the correlation, becuase of the division by m. Unfortunately, the estimates are obviously positively correlated so the bound is in fact higher. Also, because of the changing statistics of the estimator from one iteration to another, their variances are hard to estimate, we may then be content with the credibility score, c, already obtained, which reflects the extent of the variance of each partial estimator in a similar way. Hence, use

$$w_j(f_m) = \frac{c_j}{\sum_{m \in S_k} c_m} \qquad (13)$$

3.3. Initial estimation

The convergence to the correct global solution relies on a good initial estimation. Peak picking with pitch estimation can be used to find primary candidates of partials. Spurious peaks can be eliminated by comparing the height of the peak to its width and its nearest valley [1]. The multipitch estimator proposed by Klapuri in [6] is suitable for determining the fundamental frequency of each souce since it focuses on the interval between peaks and hence can cope with many pitches co-existing. Its sub-band operation also makes it robust having averaged over different subbands. Pitches obtained can be used as a guideline to organize the partials already detected into harmonic sources. To do so, the notion of harmonic concordance is adopted [7]. The measurement of harmonic distance of two frequncy compoents is given by

$$d_h(i,j) = \min \left| \log \left(\frac{f_i/f_j}{a/b} \right) \right| \qquad (14)$$

where a and b are integers within the possible range given the lowest frequency in the observed mixture. The starting references of the group are the fundamentals. If missing, other strong components may be used. The rest of the partials are then considered one by one for the minimum total harmonic distance from the partials already grouped to a particular source. If there is any ambiguity, that is, the difference of grouping a partial to one source than another is not large enough, others are grouped first and the ambigious partials will be revisited after the unambiguous ones have been assigned. This will improve the chance of grouping correctly. Also, if a harmonic of a source is missing, it is checked against the possibility of the component being assigned to the other source due to coincidence. In experiments, the initial estimation process is made sure to give reasonable estimates so that the errors do not propogate. If the partials at an expected position are perfectly overlapping, they are assigned the same parameter values for iteration.

4. SIMULATION RESULTS

A various combinations of synthetically generated harmonic sources with stationary parameters are used in the experiments at the sampling rate of 16kHz. In all experiments, T=256 and zero-padding to 1024 is employed for FFT operation prior to peak detection. The maximum number of iterations used is 100, showing relatively slow convergence compare to the algorithm in [2]. However, extension to more partials is simple and the computational complexity increases linearly with convergence gauranteed as a reward. The iteration starts on a stronger source, decided by the sum of the score of their partials, and also on a stronger partial. The algorithm shows significant improvements over the initial peak picking estimations in all cases where initialization is good.It also does better than unconstrained EM algorithm where all partials adapt freely. It is very robust to noise although a threshold effect is slightly apparent as common to many non-linear estimators as shown in Figure 1. Caution should be taken in interpreting results as curve-fitting algorithm used here should not be expected to do as well at high frequency as at low frequency while peak-picking process should be able to do equally well because of regular interval in the DFT bins. Mean absolute error is then represented with no normalization. It copes very well with coinciding and highly interfering partials. Unfortunately, best performance depends on parameter and threshold adjustment, not to mention the number of iterations allowed. The credit assignment and decision rule is

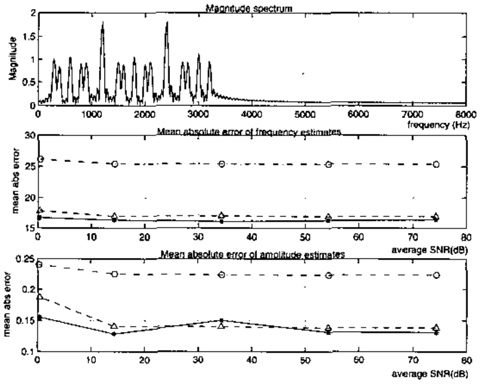

Fig. 1. Top : the mixture spectrum, Bottom two : Mean absolute error of constrained EM estimations(solid) compared to peak-picking(dash-dot) and unconstrined EM estimates(dash) in various SNR. Case of pitch=300 & 400 Hz with number of harmonics=10 & 8 respectively

verified to work in obvious cases by inspection; the coinciding or nearby partials are ruled as weak.

It should also be noted that a source with a moderate number of harmonics is more robust in estimation because of higher averaging while not too much overlapping. A stringent decision to pass a partial for strong can lead to bad results. Also, accurate amplitude estimates are much harder to obtain than the frequency, most notably when two partials are perfectly coinciding. Without any further constraint, the algorithm has no way of assigning the portion correctly. This is where the learning rate, β, can become important. Unfortunately, amplitudes of harmonic sources do not necessarily have certain relationships as in frequencies, so constraining may not be as effective. However, in this approach, none of the temporal cues available in real situation have been taken into account having considered only an estimation within a given frame of observation. The problem of coinciding partials can be further mitigated by the tracking of trajectories and the onset time can also help organize the partials into correct groups. Without temporal context, harmonically constrained EM is unlikely to yield good amplitude estimates for coinciding partials though the frequencies can still be well-constrained as shown.

To illustrate the capability in dealing with a real world signal, an example of a separation of a horn and a flute playing at different pitches is shown. The signal parameters change slightly over time as vibrato and tremolo but the estimates using a window length of about 40ms can give good estimates. Using additive synthesis reconstruction from linearly interpolated parameters across frames, closely overlapped peaks in spectrum over a sampled frame are obtained as shown in Figure 2. However, the third peak in the second source can be seen missing due to coincidence with the partial in the other source which gets all of the amplitude proportion, indicating occasional problem.

5. CONCLUSION

An alternative iterative method for sinusoidal parameter estimation of a mixture of harmonic sources is proposed. The harmonic structure allows for good estimation of the weaker partials constrained on the stronger ones based on the trustworthiness of each partial. The trustworthiness score can be calculated from the weighting of interfering spectral envelope. It is shown to give much more accurate estimates of stationary mixtures. In the future, a non-stationary model can be considered and other possible weighting scheme,maybe perceptual, could be investigated. Also, the amplitude ambiguity of coinciding partial deserves more attention and an inclusion of temporal context should also be beneficial.

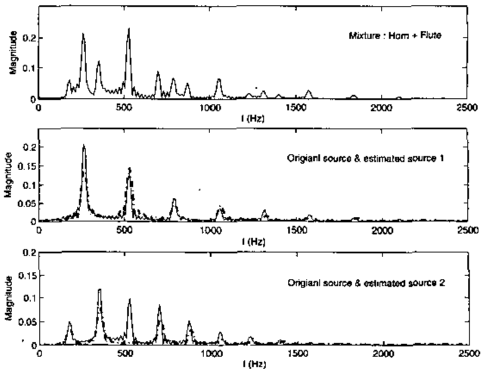

Fig. 2. From top to bottom, the spectrum of the mixture of a horn and a flute and the spectra of the original(solid) and estimated sources(dash) overlaid

6. REFERENCES

[1] X.Serra and J.O.Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition, *"Computer Music Journal"*, vol.14, no.4, pp.12-24, Winter 1990.

[2] T.Virtanen and A.Klapuri, "Separation of harmonic sounds using multiptich analysis and iterative parameter estimation," *WASPAA 2001*, 2001.

[3] M.Feder and E.Weinstein, "Parameter Estimation of Superimposed Signals Using the EM Algorithm," *Acoustics, Speech and Signal Processing*, vol.36, issue: 4, pp.477-489, 1988.

[4] N.M.Laird, A.P.Dempster and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Ann. Roy. Stat. Soc., pp.1-38, Dec. 1977.

[5] X.L. Meng and D.B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework,". *Biometrika*, 80(2):267–278, 1993.

[6] A.Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," *ICASSP2001*, vol.5, pp.9981-84, 2001.

[7] T.Virtanen and A.Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," *ICASSP2000*, vol.2, pp.765-768, 2000.

[8] A.Wang, "Instantaneous and frequency-warped techniques for source separation and signal parameterization," PhD Thesis, Stanford, 1994.

[9] S.Bregman, "Auditory scene analysis", *MIT Press*, 1990.