SUPERVISED SIGNAL PROCESSING FOR SEPARATION AND INDEPENDENT GAIN CONTROL OF DIFFERENT PERCUSSION INSTRUMENTS USING A LIMITED NUMBER OF MICROPHONES

SF Minhas School of Electrical and Electronic Engineering, University of Manchester, UK
A Barton School of Electrical and Electronic Engineering, University of Manchester, UK
P Gaydecki School of Electrical and Electronic Engineering, University of Manchester, UK

1 INTRODUCTION

This paper introduces a novel supervised offline signal processing methodology, with application to the independent separation and control of gain for different percussion instruments in a mixer.

When recordings have been made of drummers playing, it is not always clear exactly what balance of sounds will be required and thus the ability to manipulate this mix post recording is of clear benefit. Normally recordings of drummers will be made with a large number of microphones placed closely to each of the drums to allow for more recording channels to give greater control of the mix, but still each sound is not recorded in isolation and some drum sounds spill onto other channels.

The methodology covered in this paper details a way to achieve this control, with additional benefits. Using fewer microphones not only reduces cost, but in live performances can also reduce problems with feedback. Not only this, but individual drum sounds can be amplified without the amplification of other drum sounds and without compromising the integrity of the original recording.

2 SEPARATION METHODOLOGY

The motivation behind the separation methodology employed for the separation of each percussion instrument will become evident as this section progresses. The mixed signals received at the sensors from different percussion instrument can be mathematically represented by the following equation

$$x_p(n) = \sum_{q=1}^{M} \sum_{k=0}^{K} h_{pq}(k)\, s_q(n-k) \tag{1}$$

Where s_q represents the source (percussion instrument) signal that is convolved with the FIR filter containing the room impulse response (channel response) given by h_{pq} between the source and the sensor x_p. At the sensor all the convolved source signals are added to give the final convolutive mixture represented by x_p. In the above, K represents the length of the filters, M represents the total number of sources, i.e. eight in this case and n represents the sample number. Also the p represents the sensor number, in this case up to a total of three. It is an under estimated source separation case with fewer sensors than sources. Source separation for the above case can be achieved either through unsupervised or supervised adaptive filtering. However, it is a very challenging non-trivial source separation problem.

The unsupervised (blind) separation methods employ two fundamental approaches, one based on independence[1,2] and the other based on decorrelation[3]. The independence approach uses information theory and is based on the calculation of higher order statistics either explicitly[1] or implicitly[2] using a non-linear function. The fundamental objective is to exploit the density function of the originally unmixed and un-convolved signal. However, the percussion signals are discontinuous

and rhythm specific, so it is difficult to categorize them based on density function or kurtosis. A more general approach can be based on intersection of sets[3] using a slowly converging frequency domain decorrelation technique. However, all these unsupervised methodologies used in speech separation are impractical for use with percussion instrument separation, owing to many factors[4]. The most fundamental of them is the arrangement of sensors with regard to sources and room reverberation, which can make separation a mathematically ill posed problem[5] based on the inverse of the mixing matrix.

Supervised separation requires a reference signal for each percussion instrument to obtain the inverse of the mixing matrix shown by equation (1). It is not possible to play all the instruments at once and with continuation for some time till the adaptive filters converge. So, a specific rhythm based reference signal needs to be generated per percussion instrument in an anechoic (echoless) chamber. With these reference signals the separation can be achieved by playing the same specific rhythm with all percussion instruments in a real-room environment. Playing only one particular percussion instrument with a specific rhythm is quite difficult for any drummer.

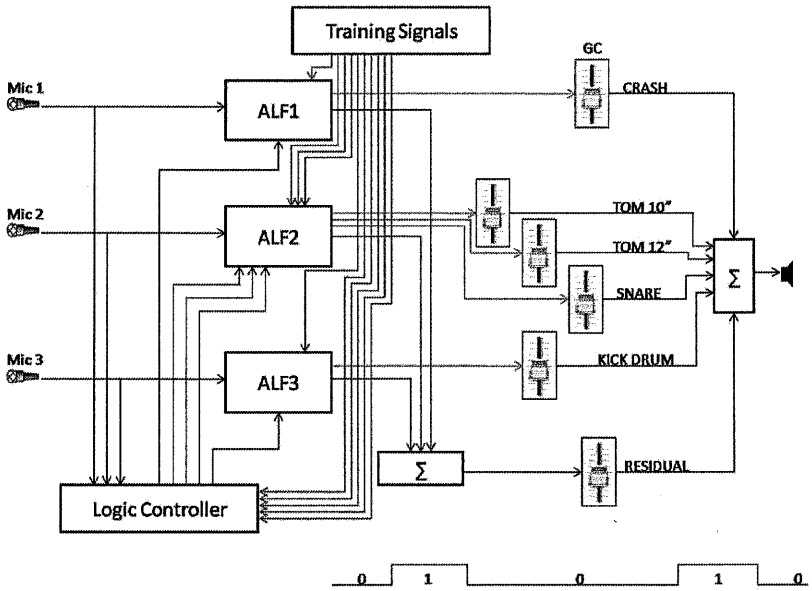

Figure 1. The flow diagram of the separation methodology

This research exploits the sporadic and the non-simultaneous nature of percussion hits and integrates a logic gate function into the mixing structure shown in equation (1) that already contains convolution and summation.

$$x_p(n) = \sum_{q=1}^{M} \sum_{k=0}^{K} h_{pq}(k)\, s_q(n-k)\, l_q(n-k) \tag{2}$$

Where $\boldsymbol{l}_q$ is a vector containing only binary values. $\boldsymbol{l}_q$ is populated with a vector of ones when a particular instrument is hit and in the duration of its percussive sound, otherwise it contains zeros. This integration simplifies the separation problem as it prevents artefact from separation occurring when it is known that a particular percussion instrument is not being played. The percussion instruments separation methodology based on the logic gate function can be seen in figure 1.

Figure 1 shows the methodology to be a two step process composed of the tracking of each instrument hit followed by an adaptive logic based filtering (ALF). The tracking of each instrument is performed in the logic controller and its output is a binary logic stream for each instrument shown by blue line that controls the gate function (switching) of ALF. The ALF is an adaptive filter based on

stochastic gradient approach and uses a least mean square (LMS) algorithm. The outputs of the ALFs, shown with green lines, are predicted separated signals and those shown with black lines are the left over residual (error) signals. The output of the ALF is then fed into the mixer, where each separated signal and residual signal can be independently controlled in magnitude using a gain controller (GC). If the all the GCs are set to 0dB, this means that no gain is applied and the final output will only be addition of the input signals from the three microphones and thus be unaffected. This separation methodology ensures that the quality of each percussion instrument is retained.

The filtering is supervised and therefore there must be a training signals block that feed a training signal (shown by red line) for each percussion instrument into the logic controller and also into the respective ALF. This training signal is distinct since it contains only one hit from each instrument in the same room environment recorded with all the microphones of the recording setup. It is typical for drummers to hit each instrument repeatedly before playing to appropriately adjust the gain at the mixer and therefore these training hits are easily accessible. Given that the training hits are recorded in the same room environment it can be assumed that the separation methodology is independent of the room environment. This makes the algorithm highly suitable to be designed as software for the post-processing of drumming signals.

2.1 Logic Controller

The logic controller is basically an intelligent tracking system for each instrument and based on that it generates a logic gate signal of **1** and **0** to control the adaptive filter of that instrument to perform separation. The bold numeric letter shows it is a vector of ones and zeros. The intelligence is based not only on the exploitation of the properties of the percussion instruments signal but also on combining the information of different instruments hit so that a hit is not misidentified. Each percussion instrument has a different sound and thus has unique spectral contents. However there are overlapping frequencies, so for this reason we will be using multiple band pass filtering for distinct frequencies followed by correlation to track the hits.

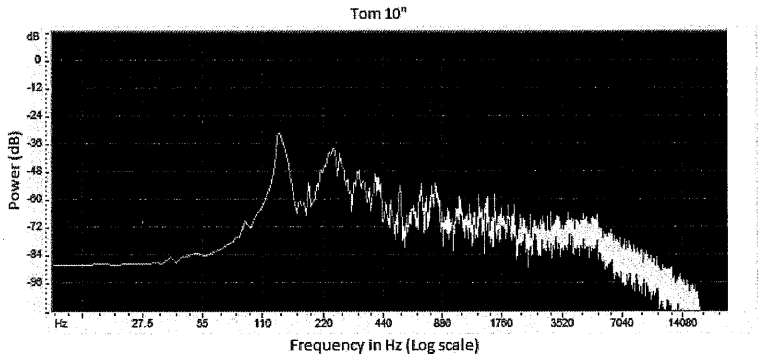

Figure 2. The frequency response of 10" tom

The typical arrangement of microphones for drum recordings is such that each instrument has at least one microphone directed at it. However, the arrangement used here has only five microphones (see analysis section), of which only three will be used. First of all the toms will be separated from the rest of the instruments followed by snare, crash cymbal and kick drum. The microphone that has the strongest toms' signal is the one that is the closest and therefore will be used for the separation of toms' signal which, in our case, is microphone 2. The spectral contents of 10" tom signal from the training hit can be seen in figure 2 with a spectral resolution of 2.929 Hz/bin at a sampling frequency f_s of 48KHz.

It can be seen from figure 2 that the 10" tom signal has strong spectral content between the bands 120 to 170 Hz and 220 to 280 Hz. These strong spectral content bands give the unique sound to 10" tom. Therefore these bands need to be exploited first before the correlation is performed. The signal from microphone 2 and also the training hit of 10" tom from the microphone 2 are passed through a very sharp cut off band pass filter with frequency response shown in figure 3.

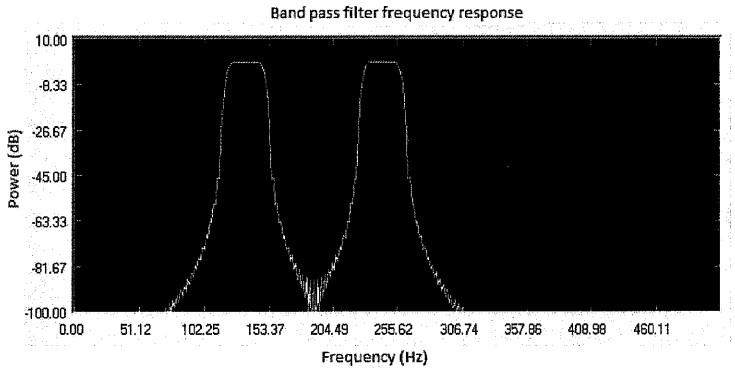

Figure 3. The dual band pass filter frequency response

The number of taps used is 16383 and the a Hanning windowing function is used. The pass bands selected are between 120 to 150 Hz and 228 to 258Hz. The filtering operation can be seen in the following equations:

$$z_2(n) = \sum_{k=0}^{K} g_1(k)\, x_2(n-k) \tag{3}$$

$$z_{2t}(n) = \sum_{k=0}^{K} g_1(k)\, x_{2t}(n-k) \tag{4}$$

Where, x_2 is the drum signal coming from microphone 2 and x_{2t} is the training signal from microphone 2, containing a single hit of 10" tom. The next step is to perform the correlation as shown by the following equation:

$$\rho_{z_2 z_{2t}}(n) = \sum_{l=0}^{L} x_2(n+l)\, x_{2t}(l) \tag{5}$$

Where L is the length of the block over which the correlation is performed. The correlation obtained from the above function needs to be normalized, so that a threshold can be adjusted to pick up the 10" tom hits. This normalization is based on the maximum energy hit throughout the length of the rhythm and shown by the following equation

$$\gamma_1(n) = {\rho_{z_2 z_{2t}}(n)} \Big/ {max(\boldsymbol{\rho}_{z_2 z_{2t}})} \tag{6}$$

In figure 4(b), it can be seen that both the toms have strong normalized correlation factor $\gamma_1(n)$. Therefore, the threshold cannot be adjusted to pick up only the 10" tom since the 12" tom also has strong spectral contents in the second frequency band of the filter i.e. 228 to 258Hz. However, this step removes any probability of the system picking up the snare and it only picks up 10" tom and in some cases 12" tom too depending upon the value of the threshold. Additionally, the other instruments are not picked up either, since either their signal strength is so low on the microphone 2, like the kick drum, or they have minimal spectral content present within the filter bands like the crash cymbal. If only the first band is selected with a new filter, g_2, and all the above steps are repeated to obtain normalized correlation factor $\gamma_2(n)$ then only the hits from 10" tom can be picked up as shown in figure 4(c) by appropriately selecting the threshold.

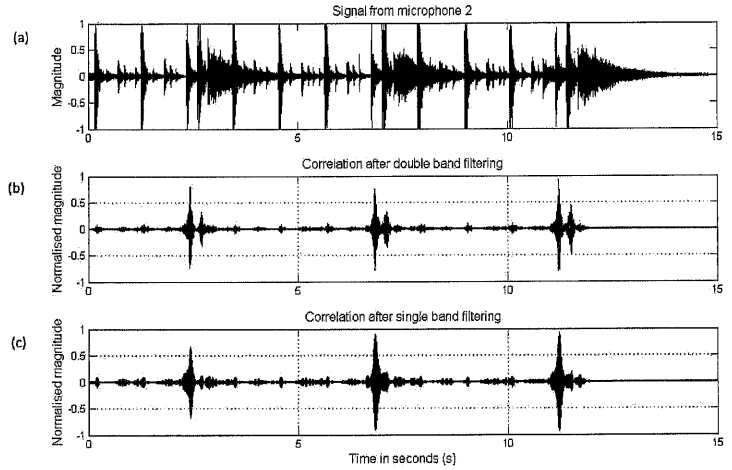

Figure 4. (a) Shows the input signal to Logic controller from microphone 1. (b) Shows the normalized correlation after double band filtering. (c) Shows the normalized correlation after single band filtering.

The 12" tom can be detected either independently or with the help of the already available hit information of 10" tom. The latter will be used due to its low computational load. The 12" tom has two strong spectral components in the bands between 70 to 110 Hz and 160 to 220 Hz. The second band has stronger spectral content than the first band, but it is common to the snare and to some extent with the crash cymbal. Therefore, passing the signal through a dual band pass filter g_3 between 80 to 110 Hz and 170 to 210 Hz and repeating the steps shown in equation 3 to 6 will remove 10" tom & crash cymbal but it will also pick up the snare in addition to the 12" tom. This specific band pass filter g_3 based correlation is avoided here but can be used in conjunction with another band pass filter to pick up 12" tom only. However, as discussed earlier, the 10" tom hit information obtained from the previous step will be used. For this reason we will not be exploiting the second band between 160 to 220 Hz for the detection. Instead, the signals will be passed through a dual band pass filter g_4 with band pass frequency ranges between 80 to 110 Hz and 228 to 258 Hz. The second band is common between 10" tom and 12" tom as discussed earlier, however the first band will ensure stronger spectral content from 12" tom. The steps shown in equation 3 to 6 are repeated with filter g_4 and both the toms will be picked up by appropriately adjusting the threshold. As the information from the 10" tom hits has already been determined, the remaining hits must therefore be from the 12" tom. Apart from the toms, the snare will also be detected from the same microphone i.e. microphone 2. For the snare, no band pass filtering is needed; only the correlation steps are performed, as shown in equation 5 and 6, using a training snare hit sample. This will also pick up the toms in addition to the snare, but with the hit information of toms already available, they can be eliminated, leaving only the snare hits.

The logic controller is designed for two fundamental functions; the first is to detect each instrument and generate a logic gate control function and the second is to share the already available hit information of different instruments. The second function ensures that individual hits cannot be identified as multiple different hits and also reduces computational load. For the kick drum the third microphone will be used, which is typically placed inside or close to a port in the kick drum. The kick drum signal is the strongest in this microphone but the toms will also be picked up since the kick drum is directly beneath and attached to them. The snare drum will also be picked up by this microphone. The kick drum hits can be detected by the correlation steps shown in equation 5 & 6, using the kick drum training signal and with the shared hit information of toms and snare.

For the detection of the crash cymbal hit, again the microphone that is closest will be used, in this case microphone 1. The crash cymbal signal contains significant energy over a long period of time when compared to other instruments, as can be seen in figure 4(a). Its spectrum consists of wideband frequencies with a slow decay rate, especially for lower frequencies. It is not possible to detect the crash cymbal using basic correlation techniques for this reason. Therefore, it is necessary to exploit its maximum energy content spread across longer period of time. For this

reason the correlation is performed after taking the root mean square (RMS) of the signal from first microphone and the training crash cymbal. Before this, however, the snare drum and toms must also be removed, otherwise the precise start of the crash cymbal sound cannot be identified. There are two ways to remove them; one way is to calculate the time delay using direction of arrival (DOA) since the toms and snare are being detected in the second microphone. The second way is to use the same values that have been saved in hit information vectors h_{t10}, h_{t12} and h_s for each of the toms and the snare respectively. The second approach is preferred as the placement of the first and second microphones is close (as seen in the analysis section, figure 5) and the maximum delay is of only a few samples. Thus the snare and toms can be closely removed by using adaptive logic based filtering. Also, the kick drum signal does not need to be removed since it is weak in microphones 1 and 2.

The final output of the logic controller is then fed into the ALF blocks to control the gate function for the separation of 10" tom,12" tom, snare , kick drum and crash cymbal are $\boldsymbol{c}_{t10}$, $\boldsymbol{c}_{t12}$, $\boldsymbol{c}_s$, $\boldsymbol{c}_{kd}$ and $\boldsymbol{c}_c$ (shown in blue lines). Here each vector $\boldsymbol{c}_{xx}$ contains a sequence of one's when each hit occurs, equivalent in length to the time it takes for the instrument sound to decay. The length of the time chosen for each instrument is different and will be discussed later on. The selection of thresholds has also not been discussed in this section and will be discussed in the analysis section.

2.2 Adaptive Logic Based Filtering

The separation of the percussion instruments is based on a novel, logic based, adaptive filtering concept. The adaptive algorithm used is least mean square (LMS) based on stochastic gradient approach using minimization of mean square error. The steps of the algorithm are shown as following

$$\hat{x}(n) = \boldsymbol{h}(n-1)\boldsymbol{y}(n) \tag{7}$$

$$e(n) = x(n) - \hat{x}(n) \tag{8}$$

$$\mathcal{J} = E\left\{\left(e(n)\right)^2\right\} \tag{9}$$

In equation (7), $\hat{x}(n)$ is the predicted signal, $\boldsymbol{h}(n-1)$ is the vector of adaptive filter coefficients and $y(n)$ is the input signal. The error term $e(n)$ is obtained by the difference of the predicted signal $\hat{x}(n)$ with the reference signal $x(n)$. The minimization of the mean square function, shown in equation (9), leads to the following filter coefficients update[6]

$$\boldsymbol{h}(n) = \boldsymbol{h}(n-1) - 2\mu\boldsymbol{y}(n)e(n) \tag{10}$$

Where μ is the step size and adjusted according to the maximum Eigen value of the correlation matrices. Different step sizes have been used for different instruments and will be discussed in the next section. This adaptive LMS filter shown by the equations (7), (8) & (10) will be used to perform separation. A control logic gate signal will be used to trigger the filtering operation. With each hit, a sequence of logic ones is generated by the logic controller to switch on the ALF. The length of the sequence is equivalent to the length of the training signal for that instrument. When no hit is detected the logic controller generates a sequence of zeros to switch off ALF, and thus no filtering is performed and the entire signal comes out as a residual. The hit information provides timing information for all the instruments. This timing information for the toms, snare, kick drum and crash cymbal instruments can be either used to pick up anomalies within a rhythm or can be used for training purpose too.

3 ANALYSIS SECTION

The arrangement of the microphones and the percussion instruments can be seen in the figure 5. The microphones are marked 1 to 5 from left to right. In figure 5, it can be seen that the first microphone is on the top of the crash cymbal and from this the crash cymbal will be removed. The second and the fourth microphones are at equidistant from the toms and snare with equivalent

signal strength and therefore only the second microphone will be considered to remove toms and snare. The third microphone is optimally placed to separate the kick drum as discussed earlier. The fifth microphone is placed to pick up the snare and Hi-hat open/closed. Since the snare has been picked up by the second microphone, the fifth microphone will also not be used. Therefore, only three microphones are used in this arrangement to separate and amplify five percussion instruments independently.

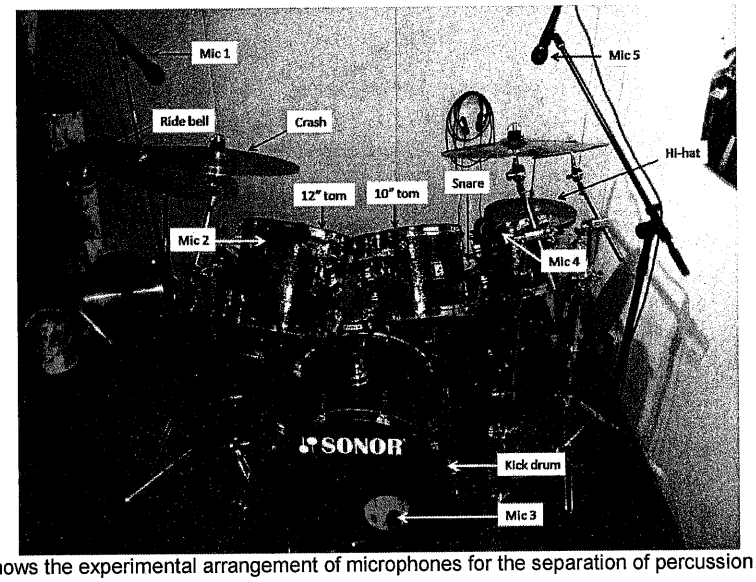

Figure 5. Shows the experimental arrangement of microphones for the separation of percussion instruments

The sampling frequency used for this experiment is 48KHz. The length of the training signal used for the 10" tom, 12" tom, snare, kick drum and crash cymbal are 1.18, 1.1, 1.12, 1.48 and 2.107 seconds respectively. The step size used for adaptive logic based separation for the 10" tom, 12" tom, snare, kick drum and crash cymbal is 0.005, 0.005, 0.0075, 0.001 and 0.005 respectively. To select the threshold the optimum range is 0.25 to 0.6 of the normalised correlation value. The separated (predicted) signals and their residual signals can be seen in the following figures.

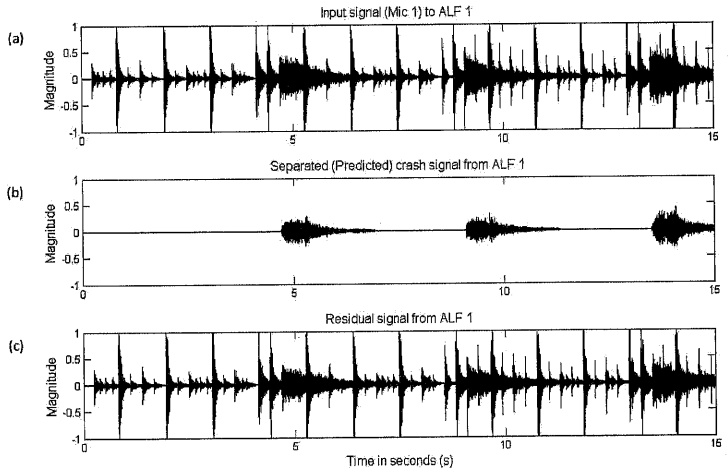

Figure 6. (a) Shows the input signal coming from the microphone 1 to ALF 1. (b) Shows the separated crash cymbal from ALF 1. (c) Shows the residual signal containing hi-hat, snare, ride bell, toms and a small part of the crash cymbal.

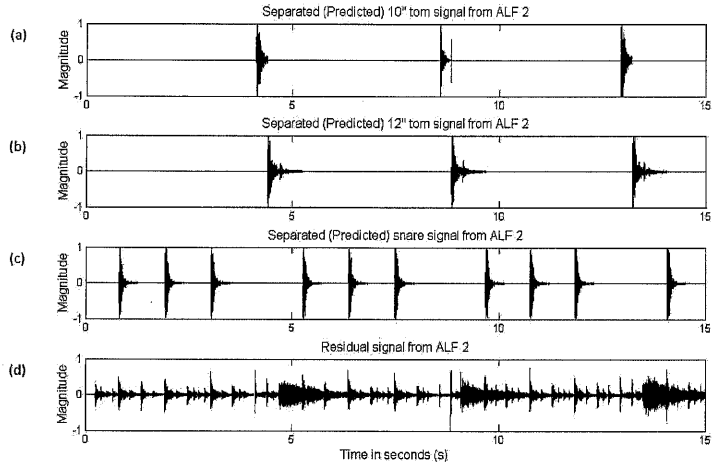

Figure 7. (a) Shows the separated 10" tom signal from ALF 2. (b) Shows the separated 12" tom signal from ALF 2. (c) Shows the separated snare signal from ALF 2. (d) Shows the residual signal containing the crash cymbal, hi-hat and ride bell.

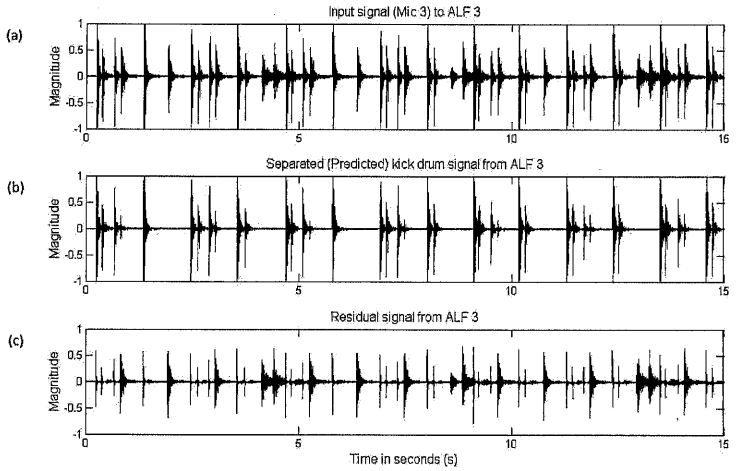

Figure 8. (a) Shows the input signal coming from the microphone 3 to ALF 3. (b) Shows the separated kick drum signal from ALF 3. (c) Shows the residual signal containing snare and toms.

4 CONCLUSION

The separation methodology presented in this paper has been shown to be able to apply gain to 5 individual percussion sounds in 3 channel drum recordings by at least ±6 dB without compromising the integrity of the drum recording. The methodology also produces timing information for the drummer allowing for assessment of the drumming and resynthesis using pre-recorded samples.

5 REFERENCES

1. A. K. Nandi, "Blind Estimation Using Higher order Statistics", 1s^{t} ed Springer, 55-56. (1999)
2. A. J. Bell and T. J. Sejnowiski, "An Information-Maximization Approacch to Blind Separation and Blind Deconvolution", MIT, Neural Computation, Vol. 7, 1129-1159. (1995)
3. D. W. E. Schobben and P. C. W Sommen, "A Frequency Domain Blind Signal Separation Method Based on De-correlation" , IEEE TSP, Vol. 50, 1855-1865. (2002)
4. S. Pedersen,J. Larsen,U. K and L. Parra, "A Survey of Convolutive Blind Source Separation Methods", Springer Handbook on Speech processing & Communication', 1-34. (2007)
5. A. Ozerov, E. Vincent and F. Bimbot, "A general flexible Framework for the Handling of Prior Information in Audio Source Separation", IEEE TASLP, Vol.20/4, 1118-1133. (2012)
6. S. Haykin and T. Kailath, "Adaptive Filter Theory", 4th ed Pearson Education,203-238 (2007)