The Use of the Attack Transient Envelope in Instrument Recognition

Benedict Tan & Deep Sen
School of Electrical Engineering & Telecommunications
University of New South Wales
Sydney, Australia

Abstract

The transient components in instrument signals have been known to contain a vast amount of information about the instrument. The undiscovered information found in the attack is known to be essential in providing the recognition of the instrument. This paper investigates the attack transient and is successful in discovering one of the features that enable the recognition of the instrument. The envelope of the attack transient has been used in this paper as a feature of the attack transient and experiments carried out showing the potential of the envelope.

1. Introduction

There are hundreds of different types of musical instruments in the world today and each of those instruments has its own characteristic that distinguishes it from another. The area of instrument recognition has been an area that has interested many researchers and engineers alike and many have been successful in being able to define features and characteristics that differentiate one instrument from another. Although there have been many features that have been found that enable the identification of certain instruments there is still so much undiscovered information waiting to be found.

Looking at the instrument signal from a temporal perspective there are four main sections, the attack, decay, sustain and retard. The main areas of focus in instrument recognition have been focused primarily on the decay and sustain portions of the music signal, these regions of the signal are also known as the steady-state. The heavy focus on the steady-state regions can be seen due to the signal being stable or pseudo-stable during those sections of the signal, because of this reason the steady-state is preferred over other sections of the signal as the data analysis is made easier due to the availability of steady-state analysis techniques. Brown has conducted studies on the steady-state using various features such as the cepstral coefficients and various statistical methods to distinguish instruments (Brown, 2001).

The other components of the instrument signal can be classified as “transient”. There are a number of studies that been conducted that show that there is a vast amount of information contained in the transient sections that enable people to recognize the instrument. The onset of the instrument signals plays a big part in characterizing the instrument and having this knowledge it is possible to use just the attack transient to be able to identify an instrument. Although the attack transient contains this information investigation into the transient is still a relatively uncovered area. The nature of the transient being non-stationary and the fact that the boundaries of the transient are not defined exactly makes it difficult to analyse. Keeler carried out experiments (Keeler, 1972) and was successfully able to differentiate between various wind instruments using temporal features such as the transient duration, delay, overshoot and the instability of the signal. Using these temporal features Keeler was able to distinguish between the different families of wind instruments. Through Keeler’s paper it was shown that even through simple perceptual features the instruments could be categorized and identified.

The purpose of this paper is to find a distinguishing feature in the attack transient that will enable the recognition of that instrument compared to another. Using the feature, systematic tests will be performed to show that the features found are feasible to be used as an attribute in instrument recognition.

In this paper the attack transient has been defined to be the non-periodic segments of the attack, which have been derived from the definition of transient, the level of harmonic content was measured and used as a gauge of the transientness of the signal. This gave a systematic way of obtaining the attack transient and also allowing

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

flexibility to be able to experiment with different levels of harmonic content in the data. Using threshold these thresholds the attack transient could be extracted from the signal and used in the following experiments.

The preliminary investigations involved analysing several attack transients of a few instruments and from the investigations revealed that there were certain perceptual attributes that were reoccurring in the attack transients. Figure 1 represents one of the attack transients from a violin with the most distinguishing features circled in. Figure 2 also shows an attack transient with the most noticeable characteristics marked. As can be seen by comparing the two figures they are not completely identical but there is a large similarity between the two samples.

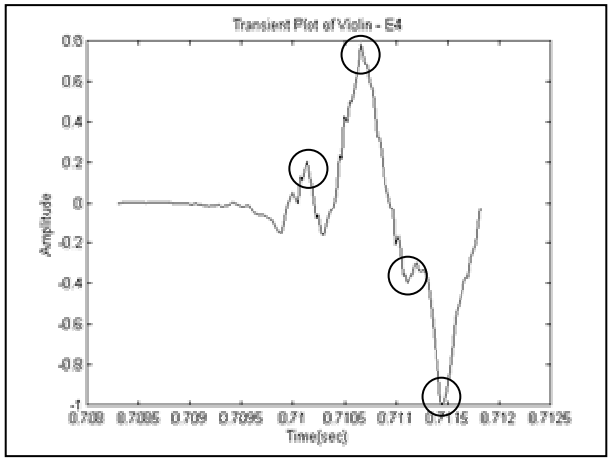
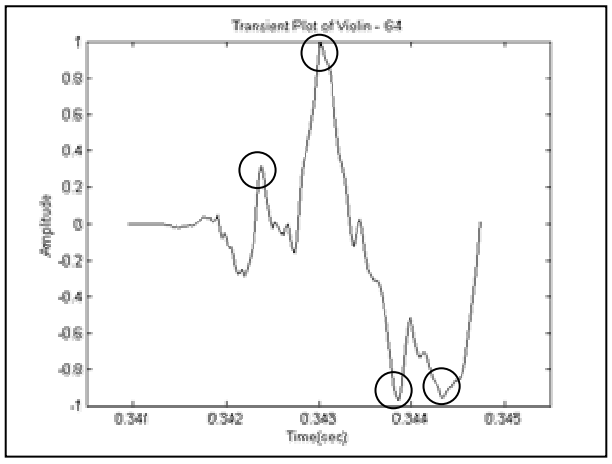


Figure 1 – Attack transient of note E4 of violin

Figure 2 – Attack transient of note G4 of violin



Further investigations into the attack transients showed that the recurring features could be pointed out in the majority of the attack transients, which lead to the possibility of being able to use the envelope as a form of identification.

The idea of using the attack transient envelope became a feasible feature for identification, but before any tests could be conducted there were a few problems that had to be dealt with concerning the use of the envelope as a feature. There was the issue of the attack transient being different lengths, although most of the features were contained in the attack transient, their duration in time varied differently and so pattern matching point to point was not feasible. A method was found which would alleviate the problem and enable pattern matching between the transient signals.

2. Dynamic Time Warping

The analysis technique chosen to analyse the instrument signals was the dynamic time warping (DTW) method. This method was primarily developed and used as a speech processing technique to be able to pattern match speech samples; it allows two speech samples that have time discrepancies to be able to be matched correctly to one another. Using this technique it is then possible to create a template of the attack transient and use the template to pattern match and compare against instrument samples to try and identify

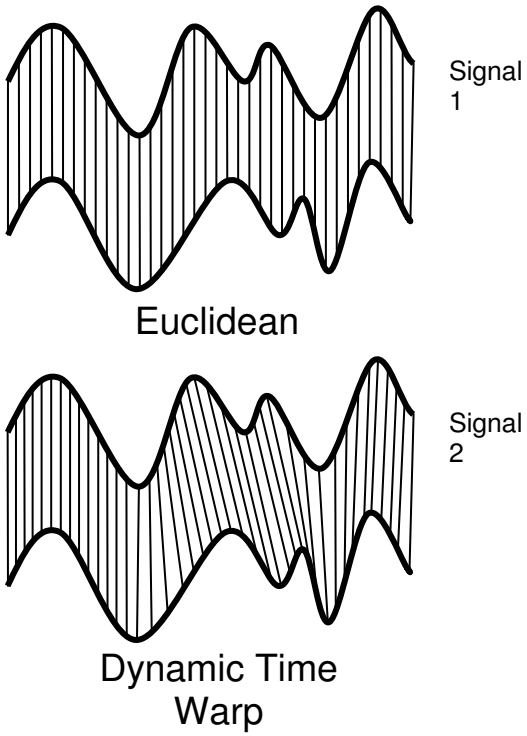


Figure 3 – Comparison of Euclidean and DTW pattern matching

the correct instrument. A comparison is shown in Fig 3 of the Euclidian based pattern matching against the DTW method. The Euclidean based pattern matching is a point to point comparison of two signals. As shown in Fig 3 the DTW is able to align the neighbouring points

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

in the sample so that the best match between the sample and template is obtained. It can be seen that through this technique the warped signal will be able to obtain a correct match to the template.

The process of how the DTW works and of which the tests were conducted is as follows. Starting with a template P and a sample signal Q to match, with length *m* and *n* respectively. We have the following

$$P = p_1, p_2, p_3, \ldots, p_m$$
$$Q = q_1, q_2, q_3, \ldots, q_n$$

An *m x n* matrix *d* is then formed of which the $d(i^{th}, j^{th})$ value being the distance between p_i and q_j element, therefore giving matrix d with the values formed by Eqn 1

$$d(i, j) = (p_i - q_j)^2 \quad (1)$$

From the local distance matrix *d* the global distance matrix *D* can then be computed. Each cell in matrix *D* is calculated by the summation of the local distance at $d(i^{th}, j^{th})$ and smallest distance of the neighbouring cells of $D(i^{th}, j^{th})$. The neighbouring cells are chosen by a stepping pattern, which will be covered in the following section. The result is matrix *D,* of which its values are the minimised global distances of the 2 sequences. A path can then be chosen by the stepping path which results in the optimum mapping of one signal to the other. The best path possible would be the straight diagonal path from the corners of the matrix which would mean that the two signals are exactly the same, therefore the more the path deviates from the optimum path of the diagonal the more distortion and warping that is needed to manipulate the sample to reflect the template.

The equation to calculate the values of matrix *D* is as follows

$$D(i, j) = d(i, j) + \min[D(i-1, j), D(i, j-1), D(i-1, j-1)] \quad (2)$$

The warping path that maps the sample to the template is the path that results in the least distortion. Starting from $D(1,1)$ the next element in the warping path will be the neighbour with the smallest value. The warping path will have a minimum length of the $\max(m,n)$ and a maximum length of $(m+n)-1$. The final value in $D(m,n)$ is the overall measure of the distortion between the 2 signals, the smaller the value the closer the match is between the template and sample and the less distortion there is in the mapping of the signals. The higher the value the more warping is needed to match the signals together.

There are certain constraints that need to be taken note of regarding the warping path such as the following:-

- Boundary Conditions: The warping path must start at $D(1,1)$ and end at $D(m,n)$.
- Continuity: the warping path can only increase by 1 point at a time; this makes sure that all points in the signal are used in the mapping.
- Monotonicity: the warping path cannot go backwards in time; this condition ensures that a point that has previously been mapped will not be mapped again.

There are various equations available to replace Eqn 2, each with different advantages and disadvantages. The algorithms chosen in this paper were the original stepping pattern shown in Eqn 2 and the Itakura algorithm in Eqn 3.

$$D(i, j) = d(i, j) + \min[D(i-1, j), D(i-1, j-1), D(i-1, j-2)] \quad (3)$$

The advantage that Eqn 3 has over Eqn 2 is that every point on the template is mapped and alleviates the problem of monotonicity. It also allows extends the range of neighbouring cells, giving the stepping pattern a larger range to be able to compare the distances. The difference in the stepping pattern can be seen in the results in the following section. Each stepping pattern has its advantages and disadvantages and there is no stepping pattern that is the overall best pattern to use. There are a number of parameters concerned when using the DTW method but it is also a powerful yet quite simple technique to use in conjunction with pattern matching and has proven to be useful in the identification of instruments as seen by the experiments.

3. Results and Discussion

The tests were conducted with two instruments, the violin and cello which come from the family of string instruments. In total here were four tests that were conducted and a total of 94 and 68 samples for the cello and violin respectively containing the third, fourth and fifth octaves. The first test conducted consisted of the stepping pattern described by Eqn 2 and the second test used the same stepping pattern with a different template for the cello instrument. The third and fourth tests were

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

the same as the previous tests but with a different stepping pattern.

The individual tests conducted were further separated into octaves and represented in the results in octaves and as a whole. Along with the results, the percentages have also been calculated showing the percentage of correct identifications for that instrument. Also included in the results following are the total for each instrument.

	cello	cello(%)	violin	violin(%)
Octave 3	7/37	19%	5/5	100%
Octave 4	13/42	31%	23/25	92%
Octave 5	4/15	27%	38/38	100%
Total	24/94	26%	66/68	97%

Table 1 – results of experiment 1

Table 1 shows the results from the first test, as can be seen the recognition rate of the violin is excellent but the results of the cello are not very good. The result from the first experiment has confirmed that the DTW is a suitable analysis technique that can be used for identification. Since the total recognition for the violin is 97%, it can be suggested that there is a bias towards the violin at this stage. However through changes in the parameters there is still a lot of room for improvement as will be seen by the following set of results.

In the next experiment the template for the cello was changed. Table 2 shows the results after the change, as can be seen there has been an improvement in the identification rate for the cello which increased 11% from the first test. The improvement in results from changing the template has shown that an improvement in the recognition rate can be obtained depending on the template chosen to represent the instrument. On the other hand an inadequate template will result in the recognition rate decreasing.

The characteristics of a good template are those that contain the various characteristics of the attack transients of that instrument. As a result it can be a rigorous testing process to find the most suitable template to represent that instrument and there might be more than one suitable candidate that is able to be used as a template.

	cello	cello(%)	violin	violin(%)
octave 3	10/37	27%	5/5	100%
octave 4	21/42	50%	23/25	92%
octave 5	4/15	27%	38/38	100%
Total	35/94	37%	66/68	97%

Table 2 – results of experiment 2

The next two experiments were executed with a changed stepping pattern which was able to improve the results even further. The stepping pattern used was that of Eqn 3 commonly described as the itakura algorithm, this stepping pattern has the advantage that the mapping of the template to the sample keeps moving forward and a point on the template can only be mapped once. This is advantageous because it is more desirable for the points on the template to be mapped once only so that a more accurate match for the instrument is obtained.

Table 3 represents the results of the test performed with the first template from the first test and also with the itakura stepping algorithm of Eqn 3. As can be seen from the results the cello recognition rate has again improved increasing a further 21%, although the recognition rate for the violin has decreased dramatically. This set of results show how the stepping pattern has great influence on the results and the ability to match the instruments. One problem with the stepping patterns is that each stepping pattern has its advantages and disadvantages and there is no best stepping pattern available. One important factor to note is that the stepping pattern chosen cannot be too stringent or too lenient. A stringent stepping pattern will result with the template only matching to the samples that are almost identical to the template and a flexible stepping pattern will allow all samples to be able to match to the template. Finding the correct median for the template is crucial to obtaining the correct results, while a stringent stepping pattern is favoured over the lenient pattern. For the experiments carried out in this paper the itakura algorithm and the original stepping pattern have been suitable in providing the results that prove that the attack envelope can be used as a feature for instrument recognition.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

	cello	cello(%)	violin	violin(%)
octave 3	17/37	46%	5/5	100%
octave 4	26/42	62%	16/25	64%
octave 5	12/15	80%	10/38	26%
Total	55/94	58%	31/68	46%

Table 3 – results of experiment 3

	cello	cello(%)	violin	violin(%)
octave 3	28/37	76%	4/5	80%
octave 4	33/42	79%	16/25	64%
octave 5	15/15	100%	24/38	63%
Total	76/94	80%	44/68	65%

Table 4 – results of experiment 4

The final test was carried out with the improved template and the itakura stepping pattern, the outcome can be seen in the table above. The results for this test have been the most improved for both instruments while increasing 22% and 19% respectively for the cello and violin. The recognition rates between the two instruments have both risen to more acceptable percentages, showing that the envelope of the attack can be used as an identifying feature for the instrument. Using the right set of parameters and templates the attack envelope can be a powerful identification feature for the instrument.

4. Conclusion

The results discussed in this paper in indicate that we have identified at least one feature in the attack transient which can be used to distinguish between musical instruments. We have shown that the feature amongst the ones investigated is the envelope of the attack transient. Of course more improvement will be possible if we looked beyond just the attack transient.

Through the use of the dynamic time warping technique the envelope of the instrument can be pattern matched to identify that instrument. By matching the features found in the attack transient it is possible to identify that instrument and the results of the experiment show that at least two instruments are able to be identified using this method.

Further work in this area will involve adding more instruments in the tests and also experimenting with more stepping patterns to increase the recognition percentages. There are a plentiful number of avenues that can be taken from this point; further research into this area will hopefully be able to improve the results further and provide a more robust way of instrument identification.

5. References

Brown, J.C (2001). “Feature dependence in the automatic identification of musical woodwind instruments”, J. Acoust. Soc. Am., Vol. 109, No. 3

Keeler, J.S (1972). “The Attack of Some Organ Pipes”, IEEE Tran on Audio and Electroacoustics, Vol. 20 no. 5 pp. 378 – 391,

Keogh, E.J and Pazzani, M.J (2001). “Derivative Dynamic Time Warping”, Department of Information and Computer Science University of California, Irvine, California USA

Tan, B (2006). “The investigation of transient components in single instrument music signals”, School of Electrical Engineering and Telecomunications UNSW, Thesis

Saldanha, E.L. and Corso, J.F. (1964) “Timbre cues and the Identification of Instruments”, Journal of the Acoustical Society of America

Wrigley S.N. “Speech Recognition by Dynamic Time Warping”, http://www.dcs.shef.ac.uk/~stu/com326/index.html

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review