Feature Extraction for Early Auditory-Visual Integration

Adrian Brown[#1], Samantha Dugelay [*2], Duncan Williams [*], Shannon Goffin [#]

[#] *Naval Systems Department, Defence Science & Technology Laboratory*
Portsdown Hill Road , Fareham, Hampshire, PO17 6AD, UK
[1] asbrown@dstl.gov.uk

[*] *Physical Sciences Department, Defence Science & Technology Laboratory*
Porton Down, Wiltshire, SP4 0JQ, UK
[2] sdugelay@dstl.gov.uk

***Abstract*—Sonar operators combine auditory and visual information to make decisions relating to target detection and identification. Many past attempts to automate the role of sonar operators have only considered the visual information and have been unsuccessful. An assessment has been made of the auditory component of the sonar operator role. This information has been used to select three algorithms with the potential for detecting features that discriminate between target types. Results are presented of the application of these algorithms to relevant time series data. Subsequently the usage of these features is discussed in terms of the concept of early auditory-visual integration.**

I. Introduction

In many applications, events are presented or displayed visually to an operator who is then responsible for detecting the presence and identity of threat targets using this visual information. This is not always effective for the detection of transient events. Such events are more likely to be detected by an auditory display and operators typically rely on listening to make a decision. This is mostly due to the human auditory system excelling in the detection of transient sounds in the presence of noise and the advantage of combined auditory and visual processing. Notwithstanding this superiority, there is still no effective way to automate this integration of auditory and visual information as part of the system display.

What we need to do is to develop a combined auditory-visual processing scheme to characterise transient events. Our starting point is to consider the ways in which submarine sonar operators interrogate auditory and visual displays. This is useful to understand some of the ways that human hearing is used alongside visual observation to separate disparate sources of transient events. In turn, this helps to provide the inspiration for some of the candidate processing methods that can be employed – this is our aim in this paper.

It is worth noting that auditory-visual processing has been developed in other applications, for example, in speech recognition [1], and while there has been success in combining audio and video features, a generalised procedure is still lacking. Different authors (e.g. [2, 3]) have advocated that the features are combined at different stages (early or late) in the processing scheme but, in general, it is first necessary to characterise (and extract) features that capture the relevant auditory and visual information.

In the following we show how features can be extracted for different sources of noise and offer comments on the way that the features could be combined. Some example results are provided to illustrate how different algorithms, that emulate the detection process, can be employed.

II. Operator Detection

Visual detection of a broadband sonar contact is determined as a line marking on a sonar trace with a bearing scale against time, marking on top of background noise. A typical example of a broadband passive sonar image is shown in Fig. 1.

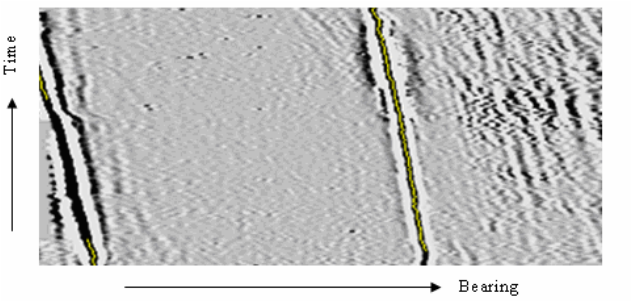

Fig. 1. Example sonar display for broadband passive sonar showing a line marking along bearing and time.

It can be observed that the display contains information that would allow an operator to choose select a direction along which to listen. This directed listening corresponds to late integration i.e. using the auditory information to confirm and/or classify the presence of a target. On the other hand, early integration would correspond to simultaneous listening and visual detection. Operators are often forced to adopt the late integration approach because different directions correspond to different beams and they can only listen to one beam at a time. The key characteristics from the auditory analysis are:

- Noise from ships consists of persistent regular rhythmic sound corresponding to motion of propeller blades;

- Noise from some marine mammals consists of a series of frequency-modulated (FM) pulses;
- Some other biological noise consists of a series of impulsive clicks.

To characterise the noise from these different sources a set of algorithms is needed that can extract rhythmic sound, FM pulses, and clicks.

III. Algorithms

This section presents an outline of the algorithms that have been selected on the basis of:

- Discrimination between types of sound;
- Performance on short samples of data.

A. *Normalised Square Different Function*

The algorithm selected to detect rhythmic sound is the Normalised Square Difference function [4]. The Square Difference Function (SDF) is defined as follows:

$$d_t(\tau) = \sum_{j=t}^{t+W-1} \left(x_j - x_{j+\tau}\right)^2 ,$$

where x is the signal, W is the window size, and τ is the lag. The SDF can be rewritten as :

$$d_t(\tau) = m_t(\tau) - 2r_t(\tau),$$

where: $m_t(\tau) = \sum_{j=t}^{t+W-1} \left(x_j^2 + x_{j+\tau}^2\right)$ and $r_t(\tau) = \sum_{j=t}^{t+W-1} x_j x_{j+\tau}$.

The Normalised SDF is then:

$$n_t(\tau) = 1 - \frac{m_t(\tau) - 2r_t(\tau)}{m_t(\tau)} = \frac{2r_t(\tau)}{m_t(\tau)} .$$

B. *Hilbert-Huang Transform*

The algorithm selected to find impulsive clicks is the Hilbert-Huang Transform [5], that is, the successive combination of the Empirical Mode Decomposition (EMD) and the Hilbert transform. EMD involves decomposing the signal into a sum of Intrinsic Mode Functions (IMFs).

The lower order IMFs capture fast oscillation modes of the signal, while the higher order IMFs capture the slow oscillation modes.

C. *Fractional Fourier Transform*

The algorithm selected to find FM pulses is the Fractional Fourier Transform [6]:

$$F^a[s(x)] = S(\omega) = \frac{\exp\left(-i\left(\frac{\pi}{4} - \frac{\alpha}{2}\right)\right)}{\sqrt{2\pi \sin\alpha}} \exp\left(-\frac{i}{2}\omega^2 \cot\alpha\right) \int \exp\left(-\frac{i}{2}x^2 \cot\alpha - \frac{ix\omega}{\sin\alpha}\right) s(x)dx$$

where $\alpha = a\pi/2$.

IV. Application to Data

Two data sets are used to illustrate the relative response of the three different algorithms:

- Marine mammal noise with frequency modulated chirps;
- Ship noise with a regular rhythm.

The approach adopted is to divide the time series data into regular "chunks" and then apply the algorithms to each chunk. The output of the algorithm can then be plotted as an output level as a function of time or frequency for each chunk.

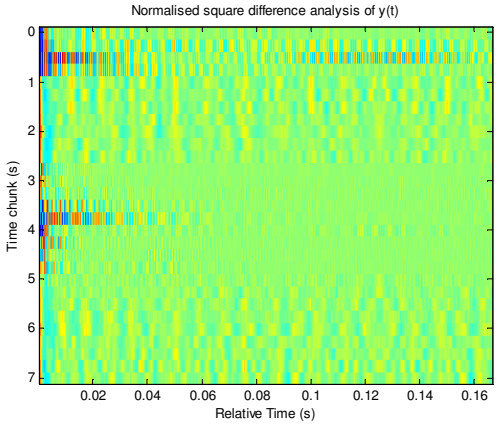

Fig. 2. Output from NSDF for marine mammal noise.

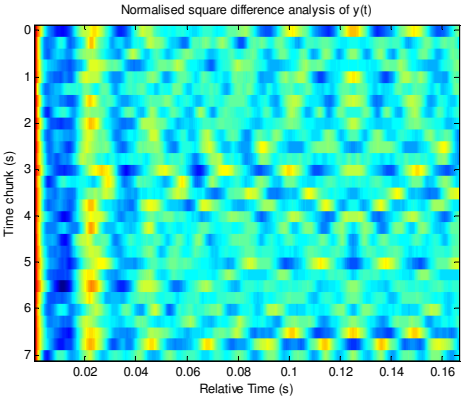

Fig. 3. Output from NSDF for ship noise.

Figs. 2—9 show the output from applying the different algorithms to each type of data.

As expected, the output from the NSDF analysis of the ship noise (Fig. 3) shows a clear persistent feature as a vertical line at 0.023 seconds corresponding to the rhythmic nature of the noise. In contrast, the NSDF analysis of marine mammal noise (Fig. 2) has no similar features.

As expected the Fractional Fourier analysis of marine mammal noise (Fig. 4) shows a clear feature as a horizontal line at 4.5 seconds. In contrast, the Fractional Fourier analysis of ship noise (Fig. 5) shows no similar features.

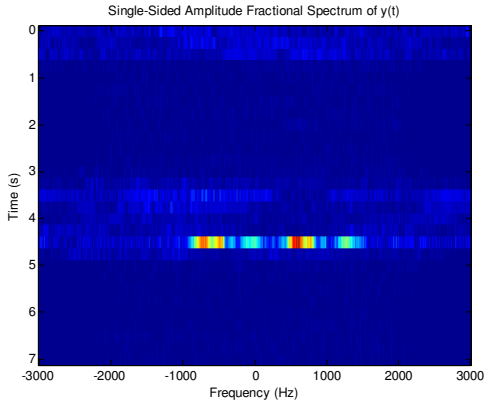

Fig. 4. Fractional Fourier analysis of marine mammal noise.

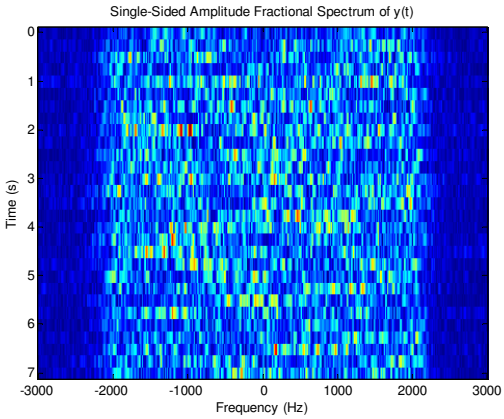

Fig. 5. Fractional Fourier analysis of ship noise.

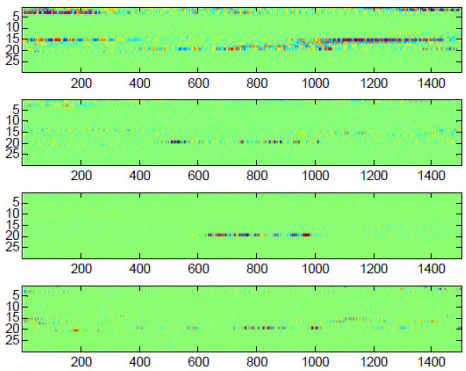

Fig. 6. IMFs of EMD of marine mammal noise.

Figs. 6 & 8 show the intrinsic mode functions (IMFs) from the Empirical Mode Decomposition (EMD) of each time chunk. In each figure the top panel is the original time series, the upper middle panel is the high frequency components with progressively lower frequency components in the lower middle and bottom panels. Figs. 7 & 9 show the Hilbert analysis of the IMFs from Figs. 6 & 8 respectively.

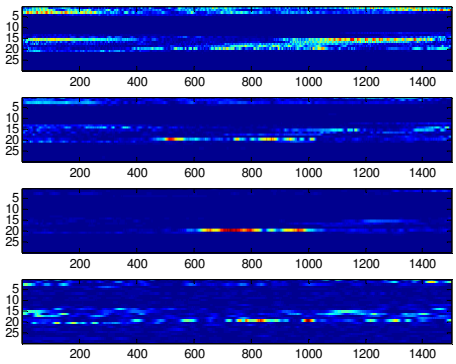

Fig. 7. Hilbert analysis of IMFs of marine mammal noise.

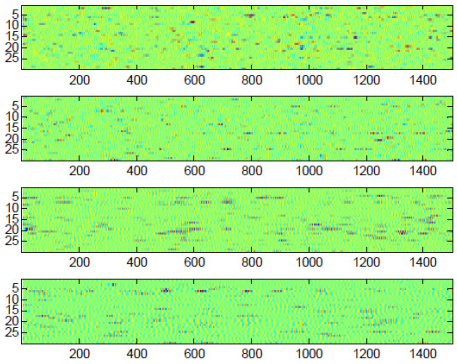

Fig. 8. IMFs of EMD of ship noise.

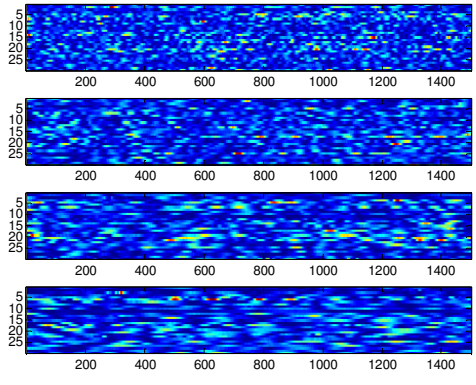

Fig. 9. Hilbert analysis of IMFs of ship noise.

The HHT analysis of marine mammal noise (Figs. 6 & 7) shows clear horizontal line features, whereas the HHT analysis of ship noise (Figs. 8 & 9) shows no similar features. Unlike the FrFT approach, the HHT algorithm does not require the pulses to have regular modulation. Hence the HHT algorithm would be expected to work against impulsive clicks as well as organised pulses.

V. Feature Extraction and Integration

A simplified version of the visual display is shown in Fig. 10 to illustrate how a feature could be extracted. Extraction of a feature from Fig. 10 could be achieved with an energy or amplitude threshold in a cell corresponding to a particular direction at a particular time.

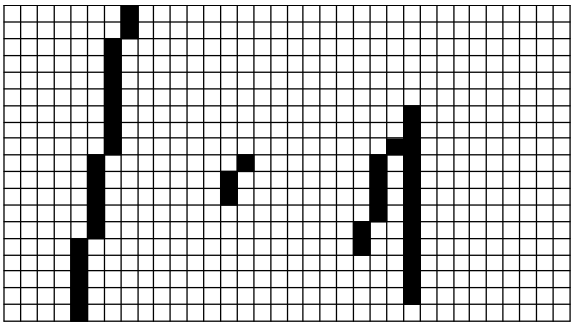

Fig. 10 Schematic view of the visual data

Plotting the output of the NSDF analysis of a rhythmic signal as a level versus time shows how quantitative information can be extracted (Fig. 11). The pitch rate, amplitude and rate of decline will vary corresponding to the periodicity and number of modes in the rhythmic signal. Hence the NSDF is able to distinguish between rhythms.

Fig. 12 illustrates feature extraction from the EMD/HHT algorithm. The Intrinsic Mode Functions (IMFs) are automatically extracted recursively until the stop criterion has been reached, i.e. no more signal fluctuations are found. The number of IMFs found can be used as a complexity feature. Each Hilbert transform of the IMFs provides a time-frequency analysis enabling the detection of transient signals.

The combination of information from the different bands (IMFs) provides the start and end time of the transient signal as well as the signal amplitude; this can provide simple features to feed into the overall audio feature vector. The spectral content of each IMF could also provide additional features such as the Hilbert coefficients.

Fig. 13 illustrates the early integration concept. Once the features from each of the algorithms are established they could be fused together into a set of joint features and used to characterise the source of noise using auditory **and** visual information. The practicality of the fusion process in this context would form a useful topic of future research.

VI. Conclusion

The potential use of algorithms to derive features in the auditory data stream that discriminate between signal types has been demonstrated. Algorithms have been found that respond to rhythmic and pulsed sound. Feature extraction from the visual representation of sonar data is known to be straightforward. Hence the remaining issue is to develop a fusion process that is effective in this context.

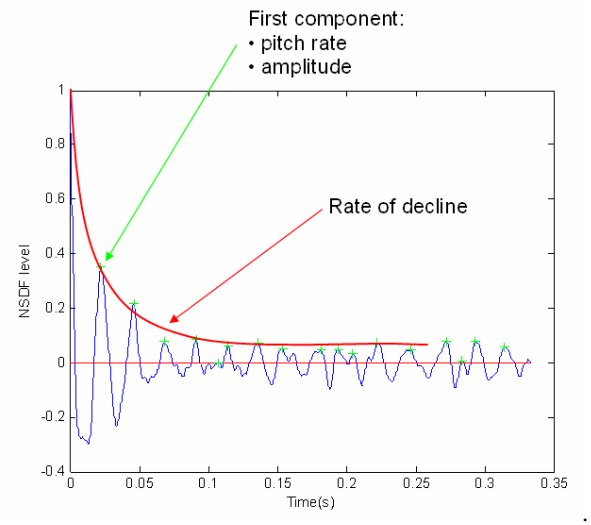

Fig. 11. Example features extracted from NSDF analysis.

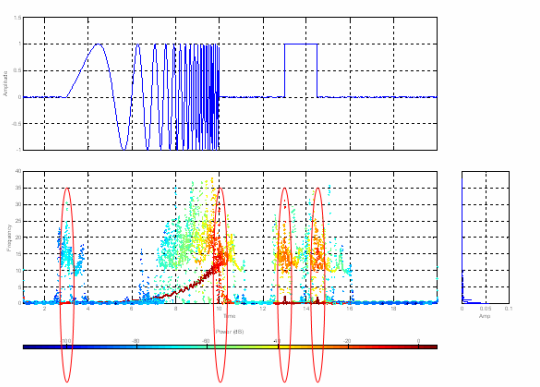

Fig. 12. Example features extracted from EMD/HHT analysis.

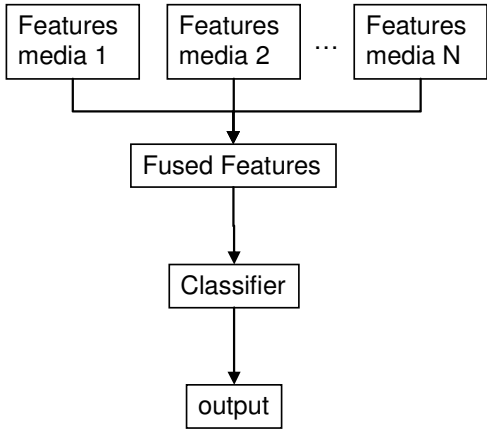

Fig. 13. Outline concept for early auditory-visual integration.

ACKNOWLEDGMENT

This work has been supported by the UK Ministry of Defence under the internal strand of the University Defence Research Centre in Signal Processing.

REFERENCES

[1] G. Potamianos, C. Neti, G. Gravier, A. Garg and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, pp 1306-1326, 2003.

[2] M. Liu and T. Huang, "Video based person authentication via audio/visual association," *Proc. ICME*, pp 553-556, 2006

[3] I. Matthews, "Features for audio-visual speech recognition", PhD thesis, School of Information Systems, University of East Anglia, Sept. 1998.

[4] P. McLeod and G. Wyvill, "A smarter way to find pitch," *Proc. International Computer Music Conference*, pp 138-141, Sept. 2005.

[5] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung and H. Liu, "The empirical mode decomposition and the Hilbert Spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. Lond. A*, pp 903-995, 1998.

[6] R. Saxena and K. Singh, "Fractional Fourier Transform: A novel tool for signal processing," *J. Indian Inst. Soc.*, pp 11-26, 2005.