2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Real time robot audition system incorporating both 3D sound source localisation and voice characterisation

Ben Rudzyn, Waleed Kadous, Claude Sammut
CAS – Centre for Autonomous Systems, School of Computer Science and Engineering,
University of New South Wales, Sydney, Australia

Abstract - **This paper describes the implementation of a novel real time robot audition system which combines a 3D sound localisation system and a voice characterisation (VC) system. The localisation system employs a 4 microphone array and uses the time delay estimation method. Accuracy is improved through the use of a correlation confidence threshold and a median filter. The VC system, which classifies between *speech*, *non speech* and *silence*, uses a decision tree classifier and a feature set comprising MFCCs, mean MFCCs and variance in MFCCs. The complete system has a processing time of 0.73x real time, and a range of up to 3 m. The compact design, high accuracy, and real time processing ability makes the system and the approach well suited to robotics.**

I. Introduction

An audition system can provide a robot with the ability to receive and process sounds arriving from any direction, without the aid of other sensory systems such as vision, thus enhancing the sensory information about its local environment. The ability to detect and direct a robot's attention towards a particular sound source has many important applications including robot navigation, particularly in object tracking or avoidance. In addition the capacity to distinguish between different incoming sounds would allow a robot to focus its attention on specific sources while ignoring others, and thus enhance human-robot interactions (HRI), and complement an automatic speech recognition system (ASR).

To effectively fulfil these applications, a robot audition system should be able to 1) identify the location of sound sources, 2) separate recorded sound waves in order to either identify multiple sources, or to isolate and focus on one particular source, and 3) extract useful information from the environment for specific applications, such as speech recognition.

This paper describes the implementation of a robot audition system known as RRAS. The inclusion of a sound localisation system fulfils the first function of an audition system, as it identifies the spatial coordinates of a sound source relative to the robot in three dimensions: azimuth (θ), elevation (φ) and distance (ρ). A voice characterisation (VC) system, which characterises the detected sound source as *human speech* or *non speech* allows the robot to focus its attention on a particular source of interest and to ignore all others, thus achieving the second primary function of an audition system. In addition, as voice characterisation is achieved through the extraction of useful information from the received waveforms, the VC system inherently fulfils the third primary function of an audition system.

Figure 1 illustrates the stages used in RRAS to process incoming sounds. Both a 3D location (relative to the robot) and a sound characterisation are produced for each frame of data recorded. A more detailed explanation of the system components follows in sections II to IV.

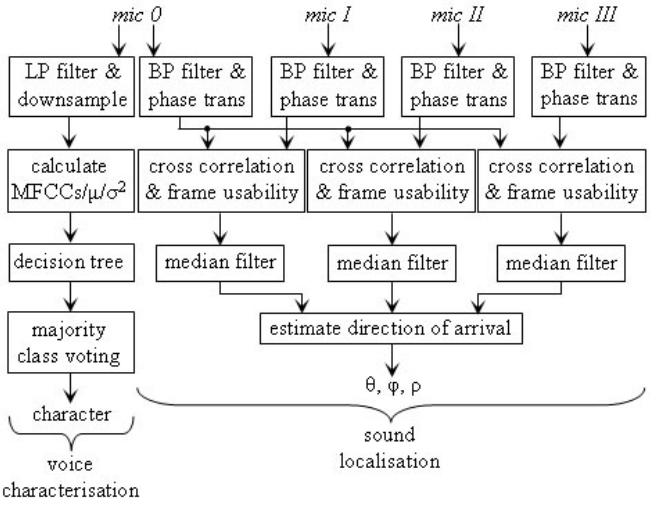

Fig 1: RRAS system layout

The approach taken here is particularly suited to robotics because the array is compact but still capable of full 3D localisation with only 4 microphones. In addition, both localisation and characterisation processing are completed in real time, largely due to the rapid testing by using a decision tree. Furthermore, the system works up to a range of 3 m, making it practical at both near and far field distances.

While localisation has previously been combined with ASR systems (e.g. [9]) to date it has not been attached to a full sound characterisation system. This is particularly beneficial to robot navigation and HRI because such a combination allows a robot to recognise and track or avoid a variety of different sources in addition to speech, such as tracking *speech* but avoiding *alarms* or vice versa.

II. Time Delay Estimation Based Sound Localisation

There are numerous examples in the literature of robot sound localisation systems which can estimate a sound source's azimuth and elevation (e.g. [12]). Similarly the development of 3D localisation systems for robots is not unique, such as that built by Bechler *et al* [1]. However, these 3D systems are either too large for practical robotic applications, or involve computationally exhaustive methods which make them difficult to implement in compact, mobile robots. In contrast, the localisation system employed in

RRAS is both computationally efficient and physically compact, thus suitable for both fixed and mobile robots.

A. *Time delay estimation (TDE)*

Given a pair of separated microphones, unless a particular sound source is equidistant from the two microphones, the propagation of a sound wave over the different path lengths will result in a delay of arrival between the two microphones. The standard technique is to assume that the distance between the two microphones is much smaller than the distance to the source so that the incident angles to the microphones can be approximated as the same (the far field approximation). With this simplification the direction of arrival (DOA) of the sound source can be calculated as:

$$\varepsilon \approx \cos^{-1}\left(\frac{d_2 - d_1}{D}\right) = \cos^{-1}\left(\frac{C\tau_{21}}{D}\right) \tag{1}$$

where C is the speed of sound in air, D is the microphone separation distance, d_1 and d_2 are the propagation distances from the source to each microphone, and τ_{21} is the time delay of arrival for the sound source between microphone 2 and 1.

The first step in the TDE method is to find the value of τ for each microphone pair, which is achieved in RRAS by using the weighted cross correlation (WCC) function [3]. WCC uses the information stored in the average magnitude difference function (AMDF) to enhance the generalised cross correlation function (GCC). The relative time delay (τ) between incoming signals is determined by finding the value of l which maximises the WCC function, where:

$$WCC(l) = \frac{GCC(l)}{AMDF(l) + \delta} = \frac{\frac{1}{N}\sum_{n=0}^{N-1} x_i(n)x_j(n+l)}{\frac{1}{N}\sum_{n=0}^{N-1}\left|x_i(n) - x_j(n+l)\right| + \delta} \tag{2}$$

(δ is a small number to prevent division by 0)

RRAS also uses a modified version of the Phase Transform (PHAT) weighting filter to enhance system performance. The PHAT weights all frequencies in the frame equally, thus relying solely on the phase components for correlation. This technique is especially useful in reverberant environments, and has become the standard for TDE based localisation. In RRAS each signal is individually pre-filtered according to Equation 3, before being cross correlated using Equation 2.

$$X_i^{PHAT}(\omega) = \frac{X_i(\omega)}{\left|X_i(\omega)\right|} \tag{3}$$

B. *The DOA algorithm*

The second step in the TDE method involves combining the time delay information with the known geometry of the microphone array to obtain a direction of arrival (DOA) estimate for the sound source. The 2D array was composed of 4 omni-directional microphones arranged into 3 subarrays to form an equilateral triangle. Each subarray had the same microphone separation distance of 12.5 cm, which was significantly smaller in physical size than that used in similar localisation systems such as [1] and [12]. The array was positioned vertically, orthogonal to the z axis (Figure 2), under the assumption that sounds could only originate from in front of the array.

The DOA algorithm used to combine the three τ values was obtained initially from [7] and modified for the particular microphone array used in this work. The problem formulation is depicted in Figure 2.

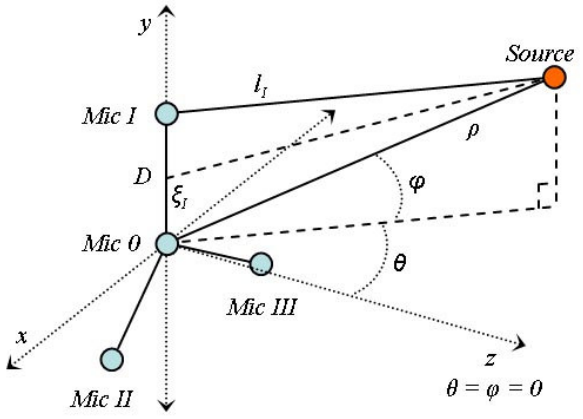

Fig 2: Geometry of microphone array and source position

For microphone 0 and I, combining the far field approximation from Equation 1 with Euclidean geometry yields Equation 4, which represents the relation of the far field approximation angle ξ_I to the near field estimates of θ, φ and ρ (D is the microphone separation distance).

$$\cos\xi_I = -\frac{D}{2\rho}\sin^2\xi_I + \sin\varphi \tag{4}$$

Similar expressions can be derived for the other independent microphone pairs. The final localisation estimation is obtained by solving these expressions, which produces:

$$\alpha = -2\frac{\cos\xi_I + \cos\xi_{II} + \cos\xi_{III}}{\sin^2\xi_I + \sin^2\xi_{II} + \sin^2\xi_{III}} \tag{5}$$

$$\beta = \frac{\cos\xi_I \sin^2\xi_{II} + \cos\xi_I \sin^2\xi_{III} - \cos\xi_{II}\sin^2\xi_I - \cos\xi_{III}\sin^2\xi_I}{\sin^2\xi_I + \sin^2\xi_{II} + \sin^2\xi_{III}} \tag{6}$$

$$\gamma = \frac{\cos\xi_I \sin^2\xi_{II} - \cos\xi_I \sin^2\xi_{III} - \cos\xi_{II}\sin^2\xi_I - 2\cos\xi_{II}\sin^2\xi_{III} + \cos\xi_{III}\sin^2\xi_I + 2\cos\xi_{III}\sin^2\xi_{II}}{\sqrt{3}\left(\sin^2\xi_I + \sin^2\xi_{II} + \sin^2\xi_{III}\right)} \tag{7}$$

where:

$$\rho = \frac{D}{\alpha},\ \sin\varphi = \beta,\ \sin\theta = \frac{\gamma}{\cos\varphi} \tag{8}$$

WCC-PHAT is first used to provide estimates for the three far field angles (ξ), which are then subsequently used to solve for the source location in terms of distance (ρ), elevation (φ) and azimuth (θ). The technique presented in this work is significantly more efficient than many of the DOA algorithms in the literature and therefore more suited to real time systems and robotics.

C. *Determination of frame usability*

In order to reduce the occurrence of mislocalised frames, RRAS employs a frame evaluation system which determines if a frame contains enough usable information for correct localisation. An examination of the recorded waveforms identified three different classes of frames: *clean* frames, *silent* frames (those with only background noise recorded), and *noisy* frames (recorded sound that contained too much noise or reverberation for accurate localisation). The frame evaluation system was designed to identify *normal* frames for localisation, and skip those that were *silent* or *noisy*, as these last two classes were the ones that produced erroneous localisations.

Figure 3 illustrates the WCC plot for a single microphone pair over all possible values of τ. It can be seen that in a *clean* frame there is one main peak in the plot, which is significantly higher than the surrounding peaks. Alternatively, in the *noisy* frame there is often two or more pronounced maxima, or no clearly dominant τ value at all (as is the case for *silent* frames). This difference in the strength of the dominant τ value was used as the criteria in the frame evaluation system.

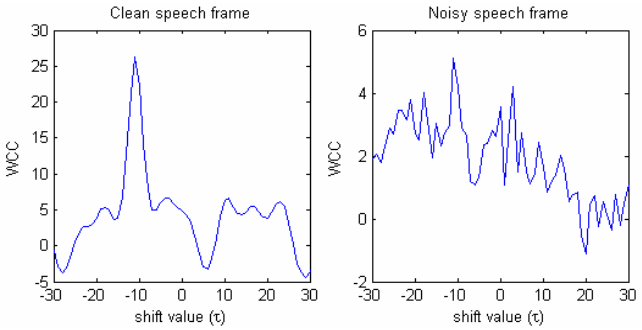

Fig 3: Comparison of correlation plots for a typical *clean* and *noisy* frame. In each case the correct shift value (τ) was -11

The height ratio between the main peak and the second highest peak for each microphone pair were summed together and compared to a threshold value (termed CCONF for correlation confidence threshold). Experimentation determined the most appropriate CCONF value to be +4.5, which was capable of correctly identifying the class of 88.1% of frames tested.

D. *Median filtering*

While the CCONF identified most of the *silent* and *noisy* frames, it was unable to exclude every mislocalised frame. Furthermore, whenever it skipped a frame there was no longer a localisation result for that frame. To compensate for this a median filter was included to help remove outliers (such as an incorrect τ estimate) and to provide localisation values for the missing frames. The filter was applied separately to each of the three τ measurements, and a new DOA estimate was calculated on the filtered values.

III. Decision Tree Based Voice Characterisation

Voice characterisation (VC) is the process of classifying a sound as *speech* or *non speech*. It is a very specialised task, and as such little work has been done in the area. However, there are a lot of similarities between voice characterisation and *speech*/*music* discrimination (SMD) which has received much more attention. Note that VC here differs from voice activity detection in that a detectable sound is not guaranteed to be speech, rather the goal is to determine what type of sound is being detected (including *silence*).

SMD researchers have examined a wide variety of different discriminators. Using a system based on Gaussian Mixture Models (GMMs) and Mel Frequency Cepstral Coefficients (MFCCs), [10] achieved a 99.5% accuracy rate for classifying *speech*. [2] also found that MFCCs and GMMs were closely suited, and obtained an overall accuracy rate of 98.8% for *speech*. Other discriminators that have been used include Hidden Markov Models [6], nearest neighbour classifiers [5] and support vector machines [8]. In general MFCCs on their own seem to provide very accurate results, regardless of the classifier used for discrimination.

However, apart from the work of [4], [2] and [5] the feature sets of choice are nearly always medium to long term (0.5 - 10 seconds) and not evaluated on a frame by frame basis (10 - 40 ms). This is acceptable in SMD, where classes such as *music* occur in long continuous blocks, but for a robot application it is essential to accurately characterise a waveform on a frame basis, so that the robot can interpret and react to multiple events in the local environment in real time.

Squires and Sammut [11] found that decision trees are well suited to speaker identification, however it does not appear that *speech/music* discrimination or voice characterisation have been explored using these learners. Decision trees would be advantageous over currently used GMMs (which provide excellent classification results) as they allow for quicker, more efficient testing.

A. *The RRAS VC system*

In a decision tree, nodes represent decision points based on attribute values, and the leaves represent the predicted class. New instances are easily and quickly classified by tracing a path down through the tree from the root to a leaf specifying the class. RRAS uses a J48 decision tree (an implementation of a pruned C4.5 decision tree) with a 0.25 confidence factor and a binary split at each leaf. The decision tree was generated using the Weka data mining toolkit [13] and then converted into *if-then-else* statements.

However, decision trees are limited in that they divide the feature space into orthogonal rectangles, where each rectangle represents a specific class. This means that for real

world data (which is rarely orthogonally distributed) decision trees can only approximate the ideal decision boundaries, inherently introducing some classification error. To compensate for this the decision was passed from the decision tree through a majority class voting system, which returned the most common class over the previous set of frames (essentially a mode filter).

The feature set of choice was the first 20 MFCCs, and the mean and variance over the last 7 frames for each of these MFCCs (i.e. 60 features in total), evaluated on a frame by frame basis (21.33 ms). The medium term duration of 7 frames (149.33 ms) was significantly shorter than that used by most researchers, such as [10] who used a 400 ms frame size and [6] who used a 1 second frame. Furthermore, medium term feature extraction was undertaken for every consecutive 7 frames, and so a decision was still made for every frame. In contrast, other researchers who incorporated medium term features used a non overlapping window, which split the test recordings into separate blocks.

Training was divided into three classification groups - *speech*, *other* (*non speech*) and *silence* - so that the system would not produce erroneous classifications when there was no sound present. In general, *silence* included minor background noise such as air conditioning or computer fans.

IV. Experimental Setup

A. Localisation

The accuracy of the localisation system was examined by recording sounds at 22 locations around the array (azimuth: 0°, +30°, +60°, +90°, elevation: 0°, ±30°, ±60°, ±90°). Due to the array's symmetry, it was assumed that readings from the left and the right hand side of the microphone array would be the same. In addition, the tests were undertaken at three distances from the array: 0.5, 1 and 3 m. Thus the system was tested at both near and far field distances in order to evaluate the validity of the far field approximation used in the DOA algorithm.

At each location, 3.2 seconds of data was recorded and processed (150 frames of 2048 samples). This was repeated for three different test sounds: a *harmonica* (representing a broadband periodic signal), a *click* (representing a broadband impulse), and *speech* (a broadband slowly varying signal). Due to the signal prewhitening caused by the phase transform, narrowband signals such as single tone sinusoids could not be localised with this system and thus all test sounds were broadband.

A positive result occurred when a frame was correctly localised inside the error threshold. A false positive occurred when a frame was deemed usable (i.e. above the CCONF threshold), but the sound was localised to a position outside the acceptable error threshold. A false negative occurred when a frame was deemed unusable, but would have been correctly localised if it had been included. The acceptable error threshold for azimuth and elevation estimation was set to 5°. There was no acceptable error threshold set for distance estimates.

Sound files were sampled at 96 kHz with a frame length of 21.33 ms in floating point. Prior to localisation, the recordings were band pass filtered at 80-16000 Hz to remove both high and low frequency background noise. This range was determined so as to retain as much useful information as possible for the three test sounds. The median filter operated over 5 consecutive frames.

All recordings contained background noise (such as computer fans and air conditioning noise), as well as being susceptible to room reverberation and multiple echoes.

B. Characterisation

150 files (50 from each class) were used in a stratified 10 fold cross validation set up for the characterisation experiment. Each file contained 149 frames of data. These files were unlabelled in that the *silent* frames in the *speech* files were still considered as *speech* for both testing and training purposes. In this way it was anticipated that the characteriser would continually characterise a speech waveform as *speech* until the speaker stopped talking, rather than alternating between *speech* and *silence* whenever the speaker paused slightly in conversation or between words. The experiment was designed to test this idea, as well as confirm that the use of medium term features could overcome the errors involved with an unlabelled system.

Characterisation was done on the same sounds recorded for the localiser, using only those from the central microphone in the array. The waveforms were first anti alias filtered at 10 kHz and then downsampled to 256 samples per frame to reduce the complexity of feature extraction. Medium term features used 7 consecutive frames, as did the voting system.

The performance of the decision tree was then compared to that of GMMs using the same files and labelling method. Three GMMs were trained (one for each class), each with 8 mixtures and a full covariance matrix. The class of the frame being tested was determined by choosing the GMM with the highest log likelihood.

V. Results

The algorithm described in Section II contains a flaw that arises whenever $\xi_I + \xi_{II} + \xi_{III} = 270°$, which led to a small α and thus a very large ρ (e.g. a distance estimate of over 100 metres). As such, the algorithm was modified by adding 0.5 to each τ estimate if this condition occurred. This modification removed all occurrences of anomalous ρ estimates, without affecting angular estimation in any way.

A. Localisation

TABLE 1
AVERAGE % ACCURACY OF THE LOCALISATION SYSTEM
(MEDIAN FILTERED RESULTS IN BRACKETS)

		Azimuth (θ)			
		0	30	60	90
Elevation (φ)	90	59.5 (55.3)	-	-	-
	60	94.8 (99.2)	95.7 (98.2)	89.0 (94.3)	62.3 (55.7)
	30	93.1 (96.8)	92.3 (97.2)	91.7 (97.4)	27.3 (7.7)
	0	95.8 (98.8)	95.3 (97.7)	94.7 (98.6)	55.7 (57.2)
	-30	97.0 (99.3)	95.7 (98.0)	91.7 (94.8)	28.7 (62.0)
	-60	97.3 (99.0)	95.5 (97.7)	84.8 (91.5)	61.7 (67.3)
	-90	5.7 (0)	-	-	-

The overall performance of the localisation system was determined by averaging the results for all 3 distances (0.5, 1, 3 m) for both the *harmonica* and *speech* recordings. Table 1 shows that the average accuracy of the system was above 84% for all angles spanning the inner ±60° (θ or φ), meaning 84% of all frames tested were accurately localised inside the error threshold. The average accuracy increased to 91% after median filtering for the same inner ±60°. There was a slight decline in performance as the distance increased away from the array, but the system was still accurate in both the near and far field cases. In fact it was more accurate in the near field case despite using the far field approximation, for example, for *speech* at 60° θ and 30° φ the accuracy rate was 96% at 0.5 m, 90% at 1 m and 83.33% at 3 m.

Performance dropped significantly outside the ±60° (θ or φ) areas due to the poor resolution of the $\cos^{-1}$ function in Equation 1 at extreme angles. A change of only 0.8 in the τ value resulted in a change of 11.2° in the estimated elevation, as illustrated in Table 2. Measurements taken in the same plane as the microphone array had the greatest chance of producing τ values close to the extremes (±34.8) and were thus most affected. The system was particularly limited in localising sounds at 90°θ 30°φ and 0°θ -90°φ due to the geometry of the microphone array, which is reflected by the very low accuracy rates from those specific locations. However, if the acceptable error threshold was increased to 15° instead of 5°, the accuracy increased for all measurements. For example those for 90° θ 0° φ increased to 88% for frame based and 96.7% for median filtered analysis, indicating that the $\cos^{-1}$ resolution was the main influence in the poor results. Of the mislocalised frames that could not be explained by this resolution, other reasons included echo interference (particularly on transitions from voiced speech to unvoiced speech or silence) or low SNR.

TABLE 2
COMPARISON OF DOA ESTIMATE OVER SMALL CHANGES IN τ

τ_{01}	τ_{02}	τ_{03}		θ	φ	ρ
34.0	-19	-19	→	0	78.8	0.79
34.8	-19	-19	→	0	90.0	0.97
34.8	-19	-20	→	-90	90.0	0.71

Echoes played a significant role in the poor performance of localising *clicks*. A *click* would last on average 1 ms, leaving another 20.33 ms for echoes to interfere with the signal. While the phase transform successfully cancelled the effect of reverberation for the *harmonica* and *speech* recordings, it was unable to do so for the much shorter *clicks*. The limiting criteria then for the above results is that the sound source must persist for at least the length of one frame (21.33 ms).

The accuracy rates in Table 1 refer to angular estimation only. Distance estimation of the system was quite poor, especially as distance increased away from the array. Table 3 displays the average % error for distance estimates with respect to azimuth and elevation. Errors in distance estimation were primarily due to the sub integer changes in shift value that were associated with a small change in source position. Equation 4 relates the far field approximation of the source direction to a near field estimation, which was particularly effective for azimuth and elevation because small changes in the incidence angle could lead to large measurable changes in the time delays (τ). However, this was not the case with distance (ρ), where large changes in ρ only resulted in small changes in τ and hence little or no change in ξ. The other limiting factor was that as the distance increased beyond 2.5 m, the time delay between sounds arriving at the microphones was below the sample rate, and thus it was not possible to register any changes in position. The only way to counter this effect would be to increase the microphone separation distance, or to increase the sample rate (which was not possible with the hardware being used).

TABLE 3
AVERAGE % DISTANCE ERROR OF THE LOCALISATION SYSTEM
(MEDIAN FILTERED RESULTS IN BRACKETS)

		Azimuth (θ)			
		0	30	60	90
Elevation (φ)	90	37.5 (34.0)	-	-	-
	60	24.5 (21.5)	30.8 (26.8)	31.8 (29.0)	63.0 (71.3)
	30	39.6 (43.3)	49.4 (46.2)	35.4 (34.4)	86.7 (87.8)
	0	54.8 (54.8)	59.7 (56.1)	103.3 (114.8)	93.1 (99.7)
	-30	41.0 (43.0)	55.3 (57.8)	82.5 (74.0)	146.0 (141.5)
	-60	49.8 (50.3)	63.8 (66.3)	65.0 (63.3)	39.0 (28.0)
	-90	67.0 (68.0)	-	-	-

B. *Characterisation*

TABLE 4
J48 AND GMM CHARACTERISATION RESULTS WITH AND WITHOUT THE VOTING SYSTEM

	Correctly characterised frames (%)			
	Speech	*Other*	*Silence*	Total
J48	97.36	95.48	98.85	97.23
J48 + voting	98.51	97.18	99.07	98.26
GMM	98.37	99.19	97.02	98.85
GMM + voting	99.08	99.42	98.24	98.91

Despite using unlabelled files, the majority of frames were still correctly characterised. This was primarily due to the use of medium term features, which caused a spreading of class information through each frame, thus the silent frames were still classified as *speech*. In addition there was most likely important information stored in the apparently *silent* sections, which helped classify them correctly as *speech* or *other* even though they appeared as *silent*.

Table 4 indicates that using GMMs for voice characterisation was slightly more accurate than using a J48 decision tree, which was mainly a result of the way that decision trees and GMMs partition the feature space for classification. However, given the same feature set and labelling method the J48 decision tree is still a viable option for use as a voice characteriser. Decision trees are advantageous over GMMs as they are easily converted into *if-then-else* statements and machine code, making then extremely fast to test and use as a classifier. In addition no *a priori* assumptions are needed about the data, and they can

easily be expanded to include more classes with minimal change in classification time.

The main drawback of decision trees though is that they are highly data dependant. Further work is needed to explore the effect of training data on the decision tree accuracy.

In all cases, the use of a simple majority class voting system effectively increased the classification accuracy with minimal change to the computational load, thus partially compensating for the inaccuracies of the decision tree classifier.

C. Real time

While the main goal of the system was to accurately localise and characterise sounds, the system needed to run in real time in order to be useful for robotic applications. Tests were run on a Pentium D 3 GHz with 2 GB of RAM, running Windows XP.

TABLE 5
RUN TIME ANALYSIS OF AUDITION SYSTEM

	Avg frame processing run time (ms)
Recording time	21.33
Localisation system [1]	15.12
VC system [2]	0.45
Complete audition system (localisation and VC)	15.57

1 - WCC (8.16 ms), CCONF (0.09 ms), BPF (1.70 ms), PHAT (4.90 ms), DOA and median (0.27 ms)
2 - MFCC calculation including LPF and downsampling (0.35 ms), decision tree and voting system (0.1 ms)

The results show that with the entire audition system being used, each frame only took an average of 15.57 ms to process a 21.33 ms frame, or 0.73x real time. This leaves 27% more processing time available within real time constraints, which could be further increased with optimisations and dedicated hardware.

The characterisation system consumes very little processing time compared to the localisation system, primarily because it only processes a single, 256 sample frame (as opposed to four, 2048 sample frames for the localiser).

D. Duration

In order to gain the best results from the system, sound sources should persist for at least 150 ms (7 frames), although the system is capable of detecting, localising and characterising sounds that are only one frame in duration. Sounds that are any shorter than this are often mislocalised due to echo interference or low SNR.

VI. CONCLUSION

This paper details the successful implementation of a novel and practical robot audition system comprising a sound localisation and a voice characterisation system. The localisation system has a minimum 91% average accuracy rate inside the ±60° range (azimuth or elevation), using a median filter and a 5° error threshold. It works well at both near and far field distances. A J48 decision tree was found to be a viable option for use as a voice characteriser, achieving a 98.26% accuracy after passing it through a voting system. The real time system is physically compact and computationally more efficient than comparable systems previously developed, making it well suited to robotic applications.

Future work involves localisation of simultaneous sources. At present if multiple sources exist, the system localises the dominant source, or alternates between sources (as is the case with multiple speakers). Distance estimation could be further improved by increasing the separation distance between the microphones.

Further work on the VC system involves the use of a more comprehensive sound database which would further test the effectiveness of the decision tree, and also allow enhancements of the system to characterise multiple sound sources (e.g. *speech*, *alarms*, *music*, *whistles*, *silence*, *other, and/or multiple speakers*).

BIBLIOGRAPHY

[1] Bechler, D, Schlosser, M and Kroschel, K, "System for Robust 3D Speaker Tracking Using Microphone Array Measurements" *Intelligent Robots and Systems*, vol. 3 pp. 2117-2122, 2004

[2] Carey, M, Parris, E and Lloyd-Thomas, H, "A comparison of features for speech/music discrimination" *International Conference on Acoustics, Speech and Signal Processing*, vol. 1 pp. 149-152, 1999

[3] Chen, J, Benesty, J and Huang, Y, "Performance of GCC and AMDF Based Time Delay Estimation in Practical Reverberant Environments" *Journal on Applied Signal Processing*, vol. 1 pp. 25-36, 2005

[4] Chou, W and Gu, L, "Robust Singing Detection in Speech/Music Discriminator Design" *International Conference on Acoustics, Speech and Signal Processing*, vol. 6 pp. 865 – 868, 2001

[5] El-Maleh, K, Klein, M, Petrucci, G and Kabal, P, "Speech/Music discrimination for multimedia applications" *International Conference on Acoustics, Speech and Signal Processing*, vol. 6 pp. 2445-2448, 2000

[6] Kim, HG and Skiora, T, "Comparison or MPEG-7 audio spectrum projection features and MFCC applied to speaker recognition, sound classification and audio segmentation" *International Conference on Acoustics, Speech and Signal Processing*, vol. 5 pp. V-925-928, 2004

[7] Lee, C, Yoon K, Lee J, and Lee, K, "Efficient algorithm for localising 3-D narrowband multiple sources" *J Radar, Sonar and Navigation*. vol. 148 no. 1 pp. 23 – 26, 2001

[8] Li, Y and Dorai, C, "SVM-based Audio Classification for Instructional Video Analysis" *International Conference on Acoustics, Speech and Signal Processing*, vol. 5 pp. V-897-900, 2004

[9] Nakadai, K, Okuno, H and Kitano, H, "Real-time Sound Source Localization and Separation for Robot Audition" International Conference on Spoken Language Processing, pp. 193-196, 2002

[10] Pinquier, J, Rouas, JL and Andre-Obrecht, R, "A fusion study in speech/music classification" *International Conference on Acoustics, Speech and Signal Processing,* vol. 2 pp. 17-20, 2003

[11] Squires, B and Sammut, C, "Automatic Speker Recognition: An Application for Machine Leaning" *International Conference on Machine Learning*, pp. 515-521, 1995

[12] Valin, J-M, Michaud, F, Rouat, J and Letourneau, D, "Robust Sound Source Localization Using a Microphone Array on a Mobile Robot" *International Conference on Intelligent Robots and Systems*, vol. 2 pp. 1228 – 1233, 2003

[13] Witten, I and Frank, E, "*Data Mining: Practical machine learning tools and techniques 2nd edition*" Morgan Kaufman, San Francisco, 2005