A Scene-Associated Training Method for Mobile Robot Speech Recognition in Multisource Reverberated Environments

Jindong Liu, *Member of IEEE*, Edward Johns and Guang-Zhong Yang, *Fellow of IEEE*
The Hamlyn Centre, Imperial College London

Abstract— **In this paper, we present a new technique for social mobile robot speech recognition based on scene-associated training models. The key contribution of the paper is a real-time framework that reduces the effect of room reverberation and ambient noise, a challenging problem in speech recognition. In classical approaches, anechoic sound is used to train the model, with the main focus on removing reverberation or noise from the sound. Our technique differs in that we train a number of speech recognizers directly from the reverberated sound, by associating each recognizer with a unique visual scene, to deal with the varying reverberation properties of different rooms. By extracting local features from a captured image and recognizing a scene, the robot can use the appropriate speech recognizer that is trained for the particular structural properties of that scene. We tested our method by using a baseline speech recognition model (HTK) across a variety of rooms and different levels of background noise. The results show that the association between a visual scene and a corresponding speech recognizer greatly improves the robot's speech recognition accuracy, together with increasing the computational speed of recognition, compared to competing techniques.**

I. Introduction

The concept of social mobile robots has been explored for many years, but most systems are still restricted to laboratory use. One of the biggest challenges is the lack of a natural human-robot communication interface. Although artificial speech recognition technology has been well developed [4], it has many limitations when transferred to a mobile robot. For example, common requirements of current speech recognition techniques demand a clean background and a close-to-microphone recording. Such requirements are difficult to satisfy when a mobile robot attempts to interpret speech at a distance, or in a noisy room, due to the effects of room reverberation and multiple sound sources [5]. In contrast, humans have a much more robust system to perform speech recognition in these complex auditory environments. The human auditory system can adapt to environment changes such as ambient noise, room reverberation differences between rooms, and even the specific location within a room. This adaptive ability in humans has inspired research into new computational auditory models to help mobile robots deal with these challenges.

One of the most well-known phenomena related to the reverberated sound perception of a human is the precedence effect [21]. Here, two spatially separated sound stimuli, with a small time delay between them, are perceived as if from a single phantom spatial position, if the inter-stimulus delay (ISD) is very small (less than 1 to 5ms). The precedence effect was explained by those neurons in the auditory system that respond to early arriving sound, subsequently suppressing the responses to later sound. Thus, a human can locate and separate sound in a reverberated environment because the reverberation arrives later than the direct sound. It was further discovered that the precedence effect exhibits adaptive behavior to the location of the sound stimuli, such as build-up and break-down [14][9]. In this case, the ISD is tuned (build-up) according to the position of the sound source and the duration of the stimulus. When the sound position changes or the reverberation environment alters, the ISD is reset (break-down). However, when the auditory scene changes back again after the ISD is reset, the previous tuned ISD returns immediately without any build-up time [15]. It is suggested that humans not only rely on the bottom-up neuron suppression to deal with reverberation, but also apply top-down memory from the cortex to change the auditory system behavior. Based on this, Blauert [7] proposed a precedence model to include a bottom-up, as well as a top-down component to take into consideration multi-model sensor inputs, such as vision. Coensel's model [13] simulated how listeners change their behavior according to attention-switch over time, based on both bottom-up and top-down cues. In addition to memory, the visual cues also play an important role for sound perception. For example, gaze direction can dynamically tune the auditory spatial map maintained by the auditory midbrain [12]. Visual attention can also override the auditory attention [30].

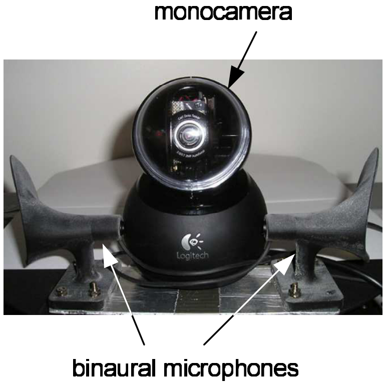

Fig. 1. Mobile robot head with binaural microphones and a monocamera.

Based on the above biological evidence, we consider that humans maintain multiple speech recognition models for various *auditory scenes*. Such auditory scenes are then associated with a *visual scene*, so that either visual or auditory cues can recall a speech recognition model when a familiar scene is revisited. In practice, we classify auditory scenes according to their impulse responses (IRs) and associate each visual scene with an auditory scene. A speech recognition model is trained for each auditory scene using reverberated training data, which is generated by convolving the IR of the scene with an anechoic sound. During operation, a robot can use vision methods to recognize a visual scene to which a unique IR signature has already been assigned. Then, the corresponding speech recognition model for the scene is employed. Our model is implemented on a mobile robot with a simple head (Figure 1) equipped with binaural microphones and a mono camera. All the auditory scene IRs are collected from real in-door environments.

The rest of this paper is organized as follows. Section II proposes a system model which includes the scene-associated speech recognition model and visual scene recognition. In Section III, experimental results are presented to demonstrate the feasibility and performance of the entire system. Detailed performance comparison of our model to a classical speech recognition system is provided. Finally, conclusions are drawn and future work is considered in Section IV.

II. System Model of Scene-Associated Speech Recognition

The proposed system includes two main stages: training and testing, as shown in Figure 2. In the training stage, we divide the experimental area into a set of rooms, each consisting of either an entire room or corridors, or a subdivision of a room based upon significant structural divisions. A number of impulse responses (IRs) are measured for each room, and then rooms are classified into different auditory scenes according to the reverberation properties of the IR. An auditory scene can be one or multiple rooms. Then, an anechoic utterance training data set is convolved with these scene-IRs to generate reverberated utterance. We extract speech features from the reverberated data and train a speech recognition model for each auditory scene. Meanwhile, a number of images are collected for each room and SURF features [6] are extracted. Visual scene models are then trained to enable subsequent scene recognition. In the test stage, the robot first captures a query image and extracts the query visual features. These features are then matched to the training database to recognize the current scene. The robot then chooses the corresponding scene-dependent speech recognition model to aid the recognition of speech detected in the room. In this section, we illustrate the details of each stage.

A. Room Impulse Response

In each room, a number of binaural room impulse responses were estimated for a number of source positions and robot locations. A mono speaker played a sweep sound as a source, and a binaural microphone set in the robot head (Figure 1) recorded the detected sound. In order to simulate a natural scenario for human-robot speech interaction, we placed the speaker and robot in common activity areas, such as a passage between tables. We avoided operation in areas of unlikely activity, such as room corners. The robot is placed at least 1 meter away from hard surfaces to avoid a local extremum of reverberation. This arrangement can measure the most representative room IR with only a minimal impact from hard surface sound reflection. The distance between the speaker and the robot was maintained between 1.5 and 6 meters, conforming to the common speech communication range for humans.

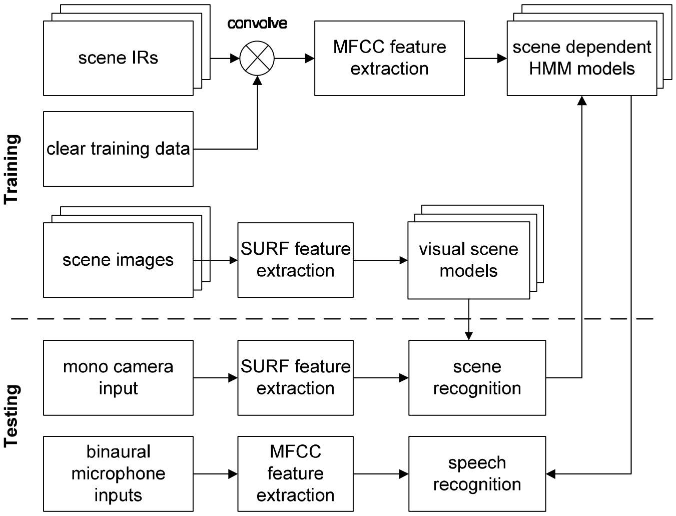

Fig. 2. Schematic structure of the scene dependent speech recognition system.

A sine sweep method [16] was applied to determine the room IR. A sweep sine wave which starts at angular frequency ω_1 and ends at ω_2 during T seconds was generated as follows:

$$s(t) = \sin\left[\omega_1 K(e^{-t/K} - 1)\right]$$

where $K = T\left(\ln\frac{\omega_2}{\omega_1}\right)^{-1}$. In our case, $T = 5$ s and the sweep wave covered 20 Hz to 20 kHz, repeating 3 times with a 5s silence between each. When the sweep signal is played through a loudspeaker and the room response recorded through microphones, the recorded signal $y(t)$ is influenced by the room reverberation, and the impulse response of the room $h(t)$ can then be calculated by:

$$h(t) = y(t) \otimes f(t)$$

where $f(t)$ is the reverse signal of $s(t)$. In another words, the convolution of $h(t)$ with $s(t)$ is an impulse signal $\delta(t)$, i.e. $f(t) \otimes s(t) = \delta(t)$. The final IR was determined by the average of the three measurements. See details in [16]. We used DSSF3 software [2] for room IR measurement. An example of a room IR is shown in Figure 3.

For the experiments conducted in this paper, we used physical constraints to define the IR measuring rooms. The physical constraints included walls, structural dividers and narrow bypasses. In each measuring room, we placed the robot at 2 positions and loudspeakers at 3 to 6 positions depending on the room size. The average of IRs at all positions was assigned to the room IR. Speech Transmission Index defined in IEC 60268-16 Third edition [18] and room acoustics parameters defined in defined in ISO3382-1 [29] were chosen to quantify room IRs. In total there are 48 parameters, including several parameters to measure the room reverberation, such as early decay time (EDT) and reverberation time T30. See details in [2]. The standardized Euclidean distances of these 48 parameters between all IRs were then calculated. If the parameter distance between two rooms was shorter than a threshold, they were merged together into an auditory scene, with the auditory scene represented by the average of IRs for each room. The threshold was selected with regards to an estimation of the minimum expected speech recognition rate after the two rooms are merged. See the experiment section for the details.

B. Speech Recognition Model Training

For speech recognition in reverberation environments, most of the current technology is focused on de-reverberation [31], i.e., the recovery of the clear sound from reverberation. The main reason for this is that the speech recognition model is traditionally trained using clear or anechoic sound. However, such methods perform poorly in real-world environments due to the effects of varying reverberation properties and unpredictable ambient noise. Additionally, de-reverberation processing requires multiple microphones. In contrast, our method brings together the reverberation and training data, allowing the speech recognition model to learn directly from reverberated data rather than clear sound. Such a method is also consistent with the human process of natural language learning as a baby, because typically our environments are echoic.

In the training module, ideally we would use training data recorded in a real scene as input. However, it is impractical to collect large quantities of training utterance for each scene. Therefore, we adopted another method to simulate the reverberated recording by convolving the scene IR with the recorded anechoic training sound. See Figure 4 for an example of a clear recording and reverberated recording. The generated reverberated sound is very similar to a real recording with regards to the extracted reverberation time parameters.

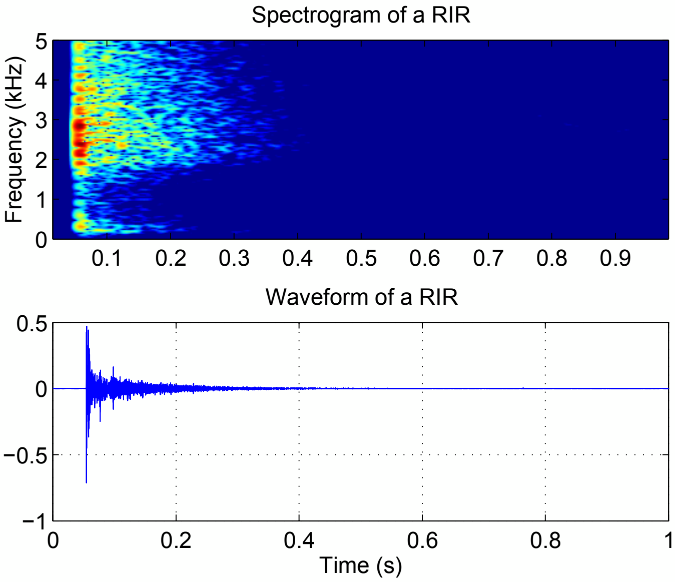

Fig. 3. Example of a measured room IR. It is the average IRs in room 306.

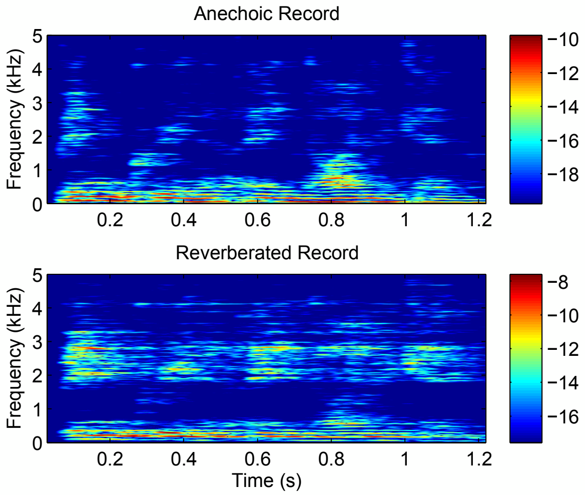

Fig. 4. The effect of room reverberation on sound recording. The recordings in both figures show speech of "bin green at A9 again". (a): anechoic recording from Grid Corpus [10], file s1-bgaa9a.wav. (b) reverberated sound after convolving with the IR of the scene 306.

Once the reverberated training data is prepared, it can be imported into any existing speech recognition training module to obtain a scene-dependent speech recognition model. In this paper, we used a general benchmark speech recognizer, HTK [26], as an example to test our system. Speech features for each training wave were extracted into standard 39-dimensional Mel frequency cepstral coefficients (MFCCs) [24], including 12 Mel-cepstral coefficients and the logarithmic frame energy, plus the corresponding delta and acceleration coefficients. Finally, all MFCC features were used to train word-level Hidden Markov Models (HMMs) with a left-to-right model topology, with no skips over states, and 32 Gaussian mixtures per state with diagonal covariance matrices [11].

C. Visual Scene Recognition

The use of visual information for mobile robot navigation has the advantage of clear geometrical information embedded in the scene. However, it has its own challenges when it comes down to practical mobile robot applications. These include illumination effects, dynamic objects in a scene, and the variation of projection of the same object from different viewpoints.

In order to deal with these issues, a popular approach is to describe images by a set of local invariant features [25] that offer improved stability over the above conditions compared to global features. Visual scene recognition then typically involves training a database of images of the environment, followed by matching a query image to the database. Matches are considered by either comparing distributions of quantized features [3], or by employing a nearest-neighbor threshold to ensure confident matches between specific features [22], followed by geometric verification [17].

Such techniques have seen success for a number of applications, including topological localization [28], loop closing in Simultaneous Localization and Mapping (SLAM) [27], and place recognition [20]. However, they rely on the assumption that a query image is captured from a near-identical location and pose to one of the database images. Whilst this is valid for applications where the robot's path can be accurately modeled and predicted, it is less practical for our case. Selecting the appropriate speech recognizer requires recognition of the surrounding scene from any arbitrary location or pose within a room. Capturing images from every feasible robot state is impractical for training, particularly in large environments, and additionally results in a slow recognition rate due to the need to attempt a match to every image in this large database.

We address this problem by adopting an approach similar to [19], whereby features are tracked across multiple training images, to form a set of *landmarks* with each representing a real-world 3D point. Each landmark $x \in \mathbf{X}$ is then assigned a mean descriptor, d_μ, from the feature track, together with the maximum descriptor distance to this mean, d_{max}, computed across all features in the track. This approach allows features in a query image to be matched to a database of landmarks, rather than to features in a database of images, enabling a more continuous representation of the environment and avoiding the somewhat arbitrary division of a room into discrete images. The number of landmarks is orders of magnitude less than the total number of features in the database, requiring significantly fewer feature match attempts. Additionally, learning the expected variance of each landmark's descriptor enables a more probabilistic approach to matching than the standard nearest-neighbor methods.

In the training stage, we extract local SURF features [6] from each training image, and track them across other images by computing descriptor distances as in [19], followed by geometric verification by estimating an affine transformation between the two images [17]. However, whilst in [19] features are only tracked between adjacent images along a tour, we attempt to track features across all images captured in the same room. This is because the path of the robot is unknown in our case, and so during training, we ensure that images are captured from a range of robot locations in an attempt to cover all possible viewpoints, resulting in the same landmark possibly re-appearing multiple times across the database. Figure 5 demonstrates landmarks tracked across a range of viewpoints and scales.

In the recognition stage, query features are compared to all landmarks in the environment, with a feature-landmark match recorded when the query feature's descriptor lies within the maximum allowed range of the landmark, based on d_μ and d_{max}. This results in a set of landmarks that are potentially present in the query image, but which may also have arisen due to false positive matches.

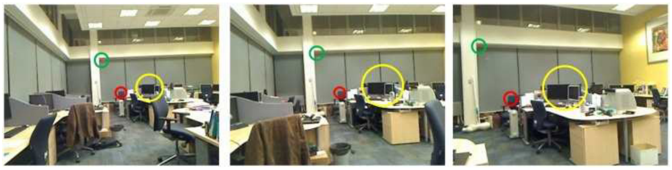

Fig. 5. SURF features detected in training images are tracked across other images in the database to form a set of landmarks. Here, all features representing the same landmark are assigned the same color.

Whilst classic approaches to scene recognition match a query image directly to database images, our database consists of landmarks independent of their original images, and so we proceed by "embedding" the query image within the pool of landmarks, as shown in Figure 6. We attempt to find the most likely location for the embedded image, by using the feature-landmark matches to assign votes to each candidate embedded location, $l \in \mathbf{L}$. Given the infinite number of possible embedded locations in the continuous environment, we assign each database landmark to the center of a new candidate location, l, denoting this landmark as the location's *embedded landmark*, x_l.

We then assign a score to each candidate location, as follows. During training, for each embedded landmark x_l representing embedded location l, a set $\mathbf{C}_l$ of co-occurring landmarks are assigned, that co-occurred in at least one image during the training stage. Then, during testing, a set of matched landmarks $\mathbf{M}_l$ is assigned to x_l, that were actually matched to a feature in the query image. The score $S(l)$ for candidate location l is then calculated by dividing the summation of the matched landmarks by the summation of the co-occurring landmarks, with each weighted by w_1 and w_2 to reflect the co-occurrence probability and landmark match probability of the landmark, respectively. As such, the score will be equal to 1 if all the landmarks that are expected to be present in candidate location l are also present in the query image.

$$S(l) = \frac{\sum\limits_{x \in \mathbf{M}_l} w_1^x w_2^x}{\sum\limits_{x \in \mathbf{C}_l} w_1^x w_2^x}$$

Weight w_1^x represents the co-occurrence probability between the matched or co-occurring landmark in $\mathbf{M}_l$ or $\mathbf{C}_l$, and the embedded landmark x_l, given that the embedded landmark is present. This is equal to the total number of training images containing both landmarks, divided by the total number of training images containing the embedded landmark. In this way, a matched/co-occurring landmark assigns a larger vote to those candidate locations whose embedded landmark co-occurs more frequently.

Weight w_2^x takes into account the discriminative power of the matched or co-occurring landmark in $\mathbf{M}_l$ or $\mathbf{C}_l$. Those landmarks with large values of d_{max} are more likely to encourage false positive feature matches, due to the greater descriptor space in which a feature can fall for a match. Thus,

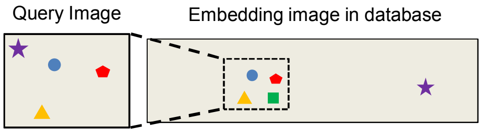

Fig. 6. Matching a set of query features to the database of landmarks requires estimating an embedding of the query image in the environment. Features in the query image on the left are matched to landmarks in the database on the right. False positive or false negative matches may occur, such as the purple star and green square, respectively, but these are filtered out by the voting strategy.

for each matched or co-occurring landmark, w_2 is assigned to the probability of a false positive feature-landmark match, computed by considering the proportion of features in the training images that are falsely matched to the landmark when comparing descriptors.

Finally, the candidate location with largest score is determined, and the room corresponding to this location is fed into the speech recognition module. Scene recognition was computed in a global sense, without prior knowledge of the robot's path. This was to ensure that the worst-case situation was dealt with, when the robot's location is highly ambiguous, such as is common in crowded social environments, or at initial start-up.

III. Experimental Results

To evaluate the performance of the proposed system, we employed a Pioneer mobile robot [1] equipped with a set of binaural microphones. A pair of external ear models were made by simulating the structure of human ears. Two cardioid microphones (Core Sound) were placed inside the ears, with 2.8 cm ear canals corresponding to the average canal length of adults [23]. One mono-camera was mounted between the two ears, with the ability to pan and tilt using build-in motors. See Figure 1 for the robot head profile. A directional mono-speaker was used as a sound source.

A. Auditory Scene Grouping

We tested our system in an environment containing a total of 15 rooms and corridors. See Figure 7. In order to find an appropriate threshold for merging rooms into auditory scenes, we conducted a preliminary test to investigate how the standardized Euclidean distance between a room's recorded IR parameters, and those IR parameters used in a training model for a different room, affect the speech recognition rate. We first randomly chose 7 rooms to represent 7 separate auditory scenes. Then we trained each room's speech recognition model and evaluated the model using the testing data (without background noise) of each of the other 6 rooms. The relationship between the standardized Euclidean distance of the IR parameters of two rooms, and the corresponding speech recognition results, is shown in Figure 8. First-order curving fitting was applied to find the correlation between distance and recognition rate (the solid line in Figure 8 shows the result). We then chose a standardized Euclidean distance threshold of 7 for grouping two rooms, resulting in an expected average recognition rate of 87%. This threshold can be adjusted based upon the specific requirements of the application, with a threshold corresponding to a higher recognition rate resulting in a larger number of auditory scenes, and hence a more time-consuming training stage.

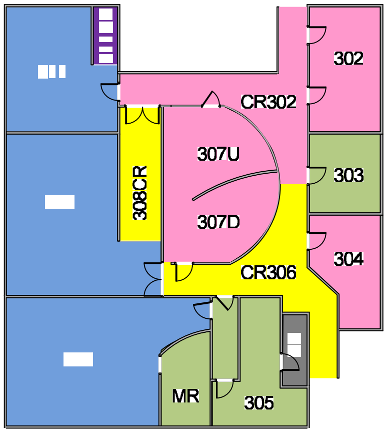

Fig. 7. Experimental room planning. CR306 and CR302 are corridors. 308CR and 311 CR are corridors in lecture rooms. All others are lecture rooms, laboratories, a kitchen and offices.

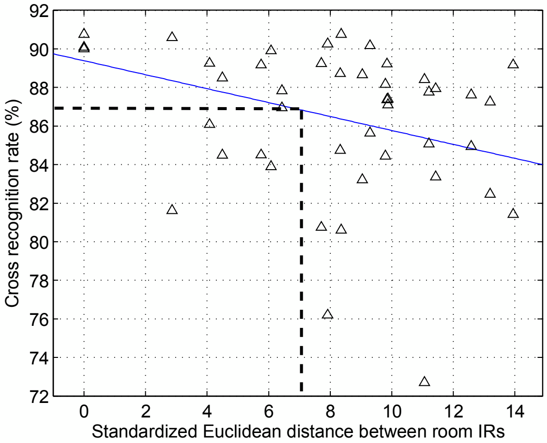

Fig. 8. The relationship between two room's IR distance and the speech recognition rate using each other's model.

Using this threshold as the maximum IR distance among one auditory scene, 15 rooms were grouped into 6 auditory scenes. This was achieved by recursively computing the two closest rooms, or groups of rooms, in parameter space, grouping them together if the distance threshold was satisfied, and updating the auditory parameters for the group as the average of the two rooms. See Figure 9 for the grouping results, where the rooms with the same scene ID are grouped into the same scene. Three IRs were recorded for each room. We found that the intra-room IR parameter distances were far smaller than the inter-room distances, and hence we naturally represented different locations within one room by the same auditory scene. Figure 7 shows the room grouping, with the same color indicating rooms are assigned to the same group, or auditory scene.

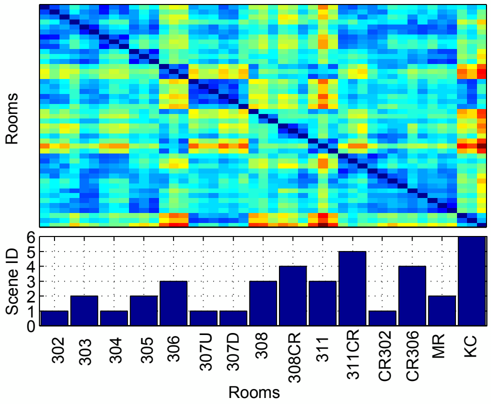

Fig. 9. The standardized Euclidean distance of room IRs and their auditory scene IDs.

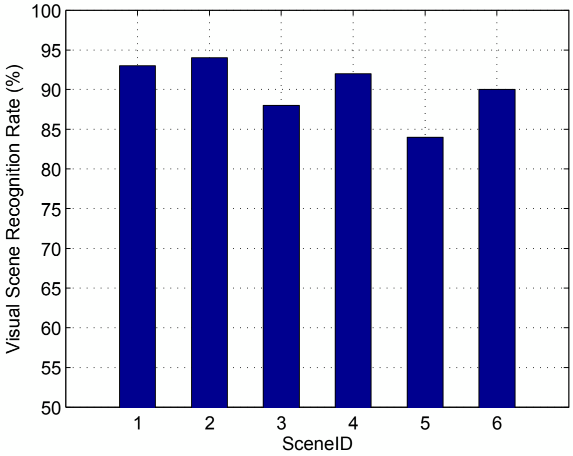

Fig. 10. The visual scene recognition results.

For speech recognition training, we adapted the Grid corpus [10] for clear training sound in Figure 2. The Grid corpus consists of utterances of 34 human speakers recorded in an anechoic room. Each utterance is a sentence with a syntax of

$$\langle command \rangle \, \langle colour \rangle \, \langle prep \rangle \, \langle letter \rangle \, \langle number \rangle \, \langle coda \rangle \quad (1)$$

Each speaker is assigned a set of words in the syntax to be read out loud. For example, Figure 4 shows an utterance of "bin green at A 9 again". We randomly selected 500 sentences from the full set of 1000 sentences for each speaker, to train each scene's speech recognizer. The remaining 500 sentences were used for evaluation. These clear sentences were convolved with each scene's IR, creating 17,000 training sentences for each scene. For comparison purposes, we also trained an anechoic speech recognition model which directly took the clear sentences as training data. The HTK model training code was partly adapted from the CHiME challenge [8].

B. Visual Scene Recognition

For the visual scene recognizer training, a number of images were captured in each room from a range of viewpoints, with the number of images per room ranging from 100 to 400 depending on the room size. For images corresponding to the same room, features were tracked across the images to generate a set of room-specific landmarks. During scene recognition, the robot captured a single image and computed the room corresponding to the image embedding with the greatest score. We tested the scene recognition performance with 50 test images per room, captured under a range of robot viewpoints. A correct match was recorded if the identified room is in the same auditory scene group as the actual room. Whilst the majority of matches arose from positive room recognition, a small number were due to a false positive match to a room that was nonetheless assigned to the same auditory scene. Figure 10 shows the recognition performance across all auditory scenes, with recognition rates ranging from 84% to 94%, at an average frame rate of 4 fps.

C. Scene-Associated Speech Recognition

Once the robot has attempted to recognize the current scene, it calls upon the corresponding speech recognizer to recognize the sound recorded in the scene. We played the utterance from the Grid corpus through the speaker to test the recognizer, generating some promising preliminary results in recognizing the spoken word. In order to evaluate our model in a systematic controlled environment, we simulated the reverberated sound in the same way as with the training data, but using a different set of clear utterances. Two sets of data were generated: i) reverberated sound without noise, and ii) reverberated sound with background noise. The first data set was obtained by convolving the clear utterances with room IRs. It was designed to evaluate our model with regards to reverberation only. However, such purely reverberated sound is rare in the practical operation of social robots because it is almost impossible to avoid background noise. As such, we generated the second testing data set by adding the reverberated sound to background noise from real-world recordings, across a range of signal-to-noise-ratios (SNRs). Here, we define the SNR as:

$$SNR = 10\log_{10}\left(\frac{\sum\limits_{t}(x_l^2(t) + x_r^2(t))}{\sum\limits_{t}(n_l^2(t) + n_r^2(t))}\right)$$

where $x(t)$ is the signal and $n(t)$ is the noise. The subscript of l and r indicate the left and right sound channel. We choose 7 SNR values, -6, -3, 0, 3, 6, 9 dB. The background noise was taken from a stereo recording in a hospital, which includes various ambient noises such as a door shutting and a baby crying.

Four speech recognition models were evaluated to compare their performance. i) The anechoic model that was trained using clear utterance only. Such a training method is applied in most of the existing speech recognition algorithms. ii) A reverb training only model, whereby the IR for the detected scene is taken as the average across all scenes. This simulates the case when the visual scene recognition is incorrect. iii) Our scene-dependent model, whereby the recognizer is chosen based upon the visual scene recognition in our experiments. iv) Our scene-dependent model in the ideal case, whereby we assume that the visual scene recognition is always perfect. The recognition rate of each model was calculated based on the correct recognition of both the letter and the number in the utterance sentence (see syntax 1). Recognition of only one of the two words was classified as unsuccessful.

Figure 11 shows the speech recognition results for reverberated sound using the four models. The anechoic model has the lowest recognition rate at an average of 15%. The three models with reverberation training significantly outperform the anechoic model, with the combined reverberation training and scene recognition model achieving 88%. Whilst our model performs better with the added scene recognition, using the reverberation training alone outperforms the anechoic model by over 60%.

Figure 12 illustrates the results when background noise was added at a SNR of 9 dB, showing a reduction in speech recognition performance across all techniques. Figure 13 shows the results of the average recognition rate over all scenes across the full range of SNR values. As can be seen, our system has a relative increase in performance as the SNR decreases, compared to the anechoic system, due to the incorporation of reverberation in the training data in our method.

The computational speed of recognition was then investigated, by averaging the times taken for recognition across all tests. This resulted in an average processing time of 1s per 3s of speech based on an Intel Core 2 Duo 2.13 GHz processor. As such, when combined with the visual scene recognition operating at 4 fps, our technique presents a framework that can run at real-time rates for practical mobile robots.

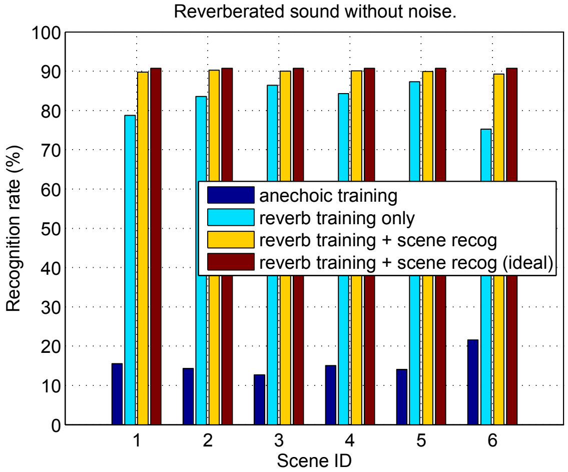

Fig. 11. Recognition results of reverberated sound without background noise.

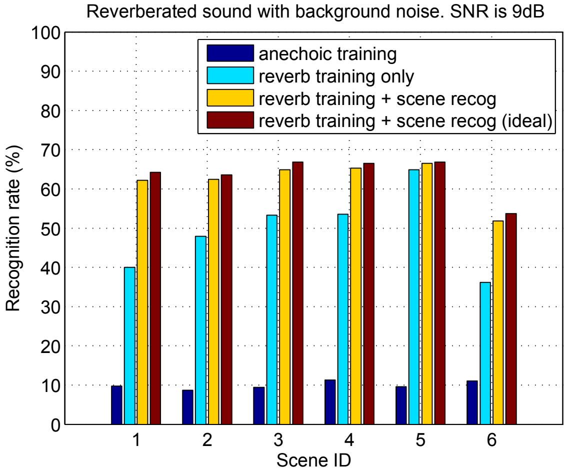

Fig. 12. Recognition results of reverberated sound with background noise. SNR = 9dB

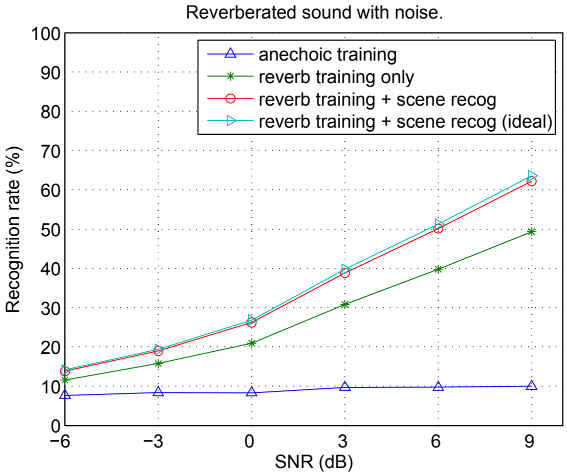

Fig. 13. Average recognition rate regarding to SNR.

IV. Conclusions and Future Work

In this paper, we have described the design and implementation of a scene-associated speech recognition method that is aimed at solving the human-robot speech interaction problem. The proposed model can deal with environments where varying reverberation behavior and multiple simultaneous sound sources can cause significant problems for classical speech recognition techniques. Our main contribution is a new system that incorporates reverberation into the training, instead of removing it, such that scene-dependent speech recognition models can be computed. A new visual scene recognition method is combined to recognize the robot's current scene from a range of viewpoints and poses, allowing the robot to use the corresponding speech recognition model that is context specific. The preliminary results shown in the paper indicate that our system can effectively deal with reverberated sound among background noise, when compared to traditional anechoic training methods.

In future work, we plan to extend our model by automating the auditory scene classification procedure, such that the robot can measure the IR of a new scene and add it to the existing set of auditory scenes. Additionally, we plan to investigate refining the auditory scene classification to provide regions with more discriminative reverberation properties, based upon a more theoretical analysis of a wider range of auditory parameters.

Acknowledgment

We would like to thanks Ning Ma and Heidi Christensen in University of Sheffield for their suggestions on HTK's training code.

References

[1] Mobile Robots Co. *http://robots.mobilerobots.com*, 2009.

[2] DSSF3 Room Acoustic Measurement. *http://www.ymec.com/products/dssf3e/*, 2011.

[3] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat. Visual topological slam and global localization. In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 4300 –4305, May 2009.

[4] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.H. Lee, N. Morgan, and D. O'haugnessy. Research developments and directions in speech recognition and understanding. Part 1. *IEEE Signal Processing Magazine*, 26(3):75–80, 2009.

[5] J. Baker, Deng Li, S. Khudanpur, Lee Chin-Hui, J. Glass, N. Morgan, and D. O'Shaughnessy. Updated minds report on speech recognition and understanding, part 2 [dsp education]. *Signal Processing Magazine, IEEE*, 26(4):78–85, 2009.

[6] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision–ECCV 2006*, pages 404–417, 2006.

[7] J. Blauert and J. Braasch. Acoustic Communication: The Precedence Effect. In *Proc. Forum Acusticum, Budapest*, pages 992–0, 2005.

[8] H. Christensen, J. Barker, N. Ma, and P.D. Green. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In *Interspeech'10*, 2010.

[9] R.K. Clifton and R.L. Freyman. The precedence effect: Beyond echo suppression. *Binaural and spatial hearing in real and virtual environments*, pages 233–256, 1997.

[10] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120:2421–2424, 2006.

[11] Martin Cooke, John R. Hershey, and Steven J. Rennie. Monaural speech separation and recognition challenge. *Computer Speech & Language*, 24(1):1–15, 2010.

[12] Qi N. Cui, Babak Razavi, William E. O'Neill, and Gary D. Paige. Perception of auditory, visual, and egocentric spatial alignment adapts differently to changes in eye position. *J Neurophysiol*, 103(2):1020–1035, February 1, 2010.

[13] B De Coensel and D Botteldooren. A model of saliency-based auditory attention to environmental sound. In *Proc. ICA,(Sydney, Australia)*, pages 1–8, 2010.

[14] T. Djelani and J. Blauert. Some new aspects of the build-up and breakdown of the Precedence Effect. *Psychological and Physiological Bases of Auditory Function*, pages 200–207, 2001.

[15] T. Djelani, J. Blauert, and A. Seestern. Modelling the direction-specific build-up of the precedence effect. In *Forum Acusticum, Sevilla, Spain*, 2002.

[16] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. *Preprints-Audio Engineering Society*, 2000.

[17] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge Univ Pr, 2003.

[18] IEC 60268-16 Third edition. Sound system equipment- part 16: Objective rating of speech intelligibility by speech transmission index. *International Electrotechnical Commission, Geneva, Switzerland (2003-05)*, 2003.

[19] E. Johns and G.-Z. Yang. Global localization in a dense continuous topological map. In *Robotics and Automation, 2011. ICRA '11. IEEE International Conference on*, page in press, May 2011.

[20] J. Kosecka and Xiaolong Yang. Global localization and relative pose estimation based on scale-invariant features. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 319 – 322 Vol.4, 2004.

[21] R.Y. Litovsky, H.S. Colburn, W.A. Yost, and S.J. Guzman. The precedence effect. *Journal of the Acoustical Society of America*, 106(4 I):1633–1654, 1999.

[22] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[23] S. Mehrgardt and V. Mellert. Transformation characteristics of the external human ear. *The Journal of the Acoustical Society of America*, 61(6):1567–1576, 1977.

[24] P. Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, 116:91–103, 1976.

[25] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005.

[26] J. Odell, D. Ollason, P. Woodland, S. Young, and J. Jansen. *The HTK book for HTK V2. 0.* Cambridge University Press, Cambridge, UK, 1995.

[27] V. Pradeep, G. Medioni, and J. Weiland. Visual loop closing using multi-resolution sift grids in metric-topological slam. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1438 –1445, 2009.

[28] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(2):300 –312, 2007.

[29] ISO Standard. 3382. Acoustics–Measurement of the reverberation time of rooms with reference to other acoustical parameters. *International Standards Organization*, 1997.

[30] Barry E. Stein and M. Alex Meredith. *The merging of the senses*. MIT Press, 1993.

[31] Mingyang Wu and DeLiang Wang. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):774–784, 2006.