AUTOMATIC MATCHED FILTER RECOVERY VIA THE AUDIO CAMERA

Adam E. O'Donovan, Ramani Duraiswami, and Dmitry N. Zotkin

Perceptual Interfaces and Reality Lab
Institute for Advanced Computer Studies (UMIACS)
University of Maryland, College Park College Park, MD 20742 USA

ABSTRACT

The sound reaching the acoustic sensor in a realistic environment contains not only the part arriving directly from the sound source but also a number of environmental reflections. The effect of those on the sound is equivalent to a convolution with the room impulse response and can be undone via deconvolution – a technique known as matched filter processing. However, the filter is usually pre-computed in advance using known room geometry and source/receiver positions, and any deviations from those cause the performance to degrade significantly. In this work, an algorithm is proposed to compute the matched filter automatically using an audio camera – a microphone array based system that provides real-time audio images (essentially plots of steered response power in various directions) of environment. Acoustic sources, as well as their significant reflections, are revealed as peaks in the audio image. The reflections are associated with sound source(s) using an acoustic similarity metric, and an approximate matched filter is computed to align the reflections in time with the direct arrival. Preliminary experimental evaluation of the method is performed. It is shown that in case of two sources the reflections are identified correctly, the time delays recovered agree well with those computed from geometric constraints, and that the output SNR improves when the reflections are added coherently to the signal obtained by beamforming directly at the source.

Index Terms— Architectural acoustics, matched filters, acoustic arrays, acoustic position measurement, deconvolution.

1. INTRODUCTION

The signal recorded at a microphone in a room incorporates the direct arrival of the sound from the source as well as the multiple weaker copies of the same signal that are created by sound reflections off the room walls. The effect of the environment on the signal can be characterized by a linear time-domain filter known as the room impulse response (RIR). The RIR length is often substantial (may be as much as a few seconds). While it can be computed using either simple geometric computations [1] [2], more advanced ray-tracing techniques [3] [4], or even numeric methods for the complicated scatterer shapes [5], the computations are expensive. Further, RIR inversion in an attempt to derive a deconvolution filter is a numerically unstable procedure [6] [7], and in order to derive useful results the RIR computation must be done with very high accuracy, which is impossible to achieve in realistic environments. Moreover, any source displacement by as much as a few centimeters requires RIR recomputation in order to keep higher frequencies coherent, and it is hard to obtain the three-dimensional position information necessary for RIR computation in realistic environment.

Another approach to enhance the desired signal in a mixture is to use spatial sound processing (beamforming) with a microphone array [8]. In a microphone array, several microphones are placed in a number of locations in space, and the signals arriving at those are filtered and summed up so that the signals originating from a desired location (e.g., a source) are amplified compared to the rest. Beamforming usually assumes that the location of interest is given and requires recomputation of the filters with the change of the location; some approaches also adaptively modify the filters in order to suppress unwanted interference, where the interference is broadly defined as anything uncorrelated with the source signal [9]. This is obviously ineffective in removing the reverberant parts of the signal [10]. Note that in case of microphone array the RIR is specific for each microphone in the array.

A combination of beamforming and RIR deconvolution is known as matched filter array (MFA) processing [11]. MFA can be thought as beamforming aimed not only at the sound source itself but also at its reflections. To do the MFA processing, knowledge of RIR for each microphone in the array is also necessary; it can be either computed analytically using a room model and source/receiver positions or measured in the actual environment where beamforming has to be applied. An MFA analog of simple delay-and-sum beamforming is obtained by truncating and inverting the RIR and inserting fixed time delay to make the resulting filter causal. In a simulated multi-path environment, simulations of MFA shows the SNR of the beamformer remaining independent of number of propagation paths, as these are compensated automatically by the inverse filtering. However, accurate knowledge of RIR is still necessary for the processing, and MFA performance degrades quickly with RIR inaccuracies caused by uncertainty in source position [12].

Due to complications associated with exact tracking of the target in three-dimensional environment, in many applications such as source localization and speech recognition the reverberative patterns imposed by the environment are seen as undesirable. However, MFA processing is one example of how the reverberation can be used constructively. Another approach studied in the past uses subspace methods to recover the signal [13] [14]; these algorithms tend to be quite expensive computationally. In the current work, a method is proposed to identify the significant reflections automatically; thus, an approximation of the inverse RIR is computed on the fly. More specifically, an "audio camera" [15] (a 64-channel spherical microphone array) is used to compute the "audio image" – a map of acoustic energy distribution in the space surrounding the array, similarly to what is done in steered response power (SRP) beamforming [16]. The acoustic sources, *as well as their significant reflections*, manifest themselves as peaks in the audio image. These peaks are found using gradient ascent and are grouped into several sets based on acoustic similarity measure. Time delays be-

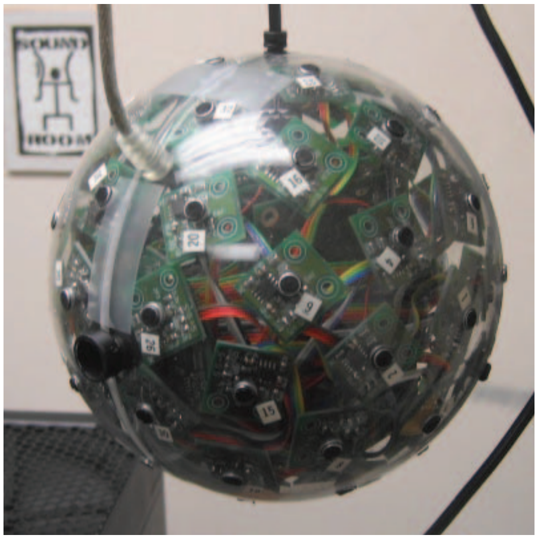

Fig. 1. An audio camera used for acoustic imaging – a 64-channel spherical microphone array with a video camera embedded.

tween signals in each group are found, and beamformer outputs from the identified reflective directions are delayed appropriately and are summed up coherently. The algorithm was tested on sample two-speaker data obtained in a conference room using speech from the TIMIT database [17], and the results were found to be consistent with expectations.

2. AUDIO IMAGING

The audio camera used for this work is shown in Figure 1. It is a 64-channel spherical microphone array with a video camera built in. The symmetric configuration allows for digital steering of the beampattern in arbitrary direction without distorting the beampattern shape. An elegant beamformer for the spherical array framework was first described in [18] and then refined by many (e.g., [19] [20]). The beamformer operates by decomposing the acoustic wavefield into a number of spherical modes corresponding to elementary solutions of the acoustic wave propagation equation and then combining the modes with weights derived from the desired beampattern. In the current implementation, high-resolution real-time audio imaging of the space is achieved by operating a number of independent beamformers in parallel on graphical processing hardware (GPU). By beamforming in many directions and plotting the energy in the output signal as a function of direction, an "audio image" is created. Sample audio image is shown in Figure 2. The projection equations of the audio camera are essentially the same as for common (video) central projection camera [21], which allows one to use many algorithms developed for computer vision (such as multi-camera calibration or epipolar constraints) with audio camera or with multiple modalities. In particular, visual analysis of acoustic energy distribution in the space can be performed by calibrating audio and video cameras in the common reference frame and overlapping evolving audio and visual images of the environment [22].

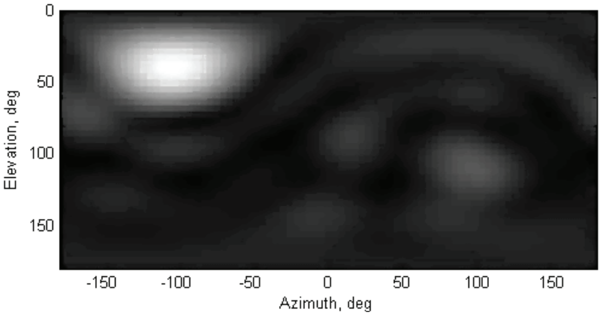

Fig. 2. Sample audio image. A source can be seen at approximately (40, -100) (elevation, azimuth).

3. CLASSIFICATION OF IMAGE PEAKS

Using the audio camera, the acoustic image is formed. Overlaid audio-video images are shown to the user, and he/she selects the target of interest on the image with the graphical user interface. A gradient ascent is run on the image to identify peaks. The algorithm is similar to the mean-shift algorithm. For greater accuracy, subpixel estimates of the true peak locations are found by interpolation near the peak position. A list of peaks in the image is then created, and the peak that is closest to the target location is assumed to be the true source location corresponding to the direct sound arrival.

The time-domain signals corresponding to all the peaks in the audio image are then computed by applying the standard beamforming procedure [18] to the directions of the peaks. These can correspond either to the reflections of the desired source, to other sources, or to their reflections. In order to determine their origin correctly, an acoustic similarity measure is used. Various similarity functions can be considered; in particular, if the sound expected is speech, the task becomes similar to the speaker verification, for which the standard mel-frequency cepstral coefficients (MFCC) are often a good choice [23]. MFCC are computed by taking a short-time Fourier transform of the signal, summing up the power in logarithmically-spaced frequency bands, and then taking a discrete cosine transform of the power values. Thirteen MFCC coefficients are used for classification. The coefficients are normalized individually across the time dimension and are fed to the vector quantization (VQ) training procedure [24] with the codebook size of 16 to build the codebook for the target (direct) signal (i.e., the signal computed by applying the beamformer to the true source location). The VQ coding error is then evaluated for signals corresponding to all the peaks in the energy map, yielding a matching score. Thresholding on the matching score is used to determine if the signal is sufficiently similar to the desired one. Signals that are similar enough are assumed to be the reflections of the desired source.

4. OUTPUT SIGNAL COMPUTATION

After identifying the reflections originating from the desired source, the algorithm proceeds with coherent summation of the signals corresponding to those, similarly to the matched filter processing. In order to do that, it is necessary to estimate the time delays between the direct signal and the reflections. One possible approach is to compute generalized cross-correlation function; however, it was found to be unreliable in reverberant conditions. Another time-delay estimation algorithm that is more robust but also more computationally ex-

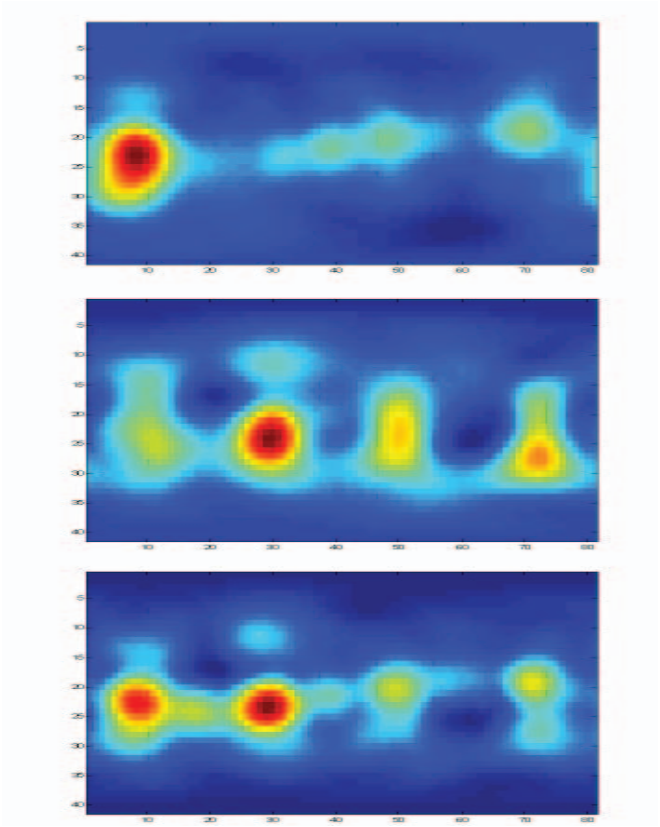

Fig. 3. Audio images for the single male speaker (top), single female speaker (middle), and both speakers active at the same time (bottom). Note that the bottom image structure is the combination of the top and the middle images.

pensive is based on adaptive eigenvalue decomposition (AED) [25]. A modified AED algorithm was used for time delay estimation in this evaluation. Once the time delays are found, the reflection signals are advanced appropriately and are added to the direct signal.

5. PRELIMINARY EXPERIMENTAL EVALUATION

We will illustrate the steps in our algorithm in the context of a two-speaker example that was acquired in a conference room. The conference room dimensions were approximately 6.4 m × 4 m × 2.4 m. A large oval table dominated the center of the room. The side walls were fairly reverberant. The floor was treated with carpet. The ceiling was covered with standard ceiling tiles and lighting fixtures.

Two loudspeakers were placed on tripods in the room at the opposite ends of the conference table. Two sample sentences (one by male speaker and another by female speaker) from the TIMIT database were played via the loudspeakers. Figure 3 shows the audio images obtained in three scenarios: single male speaker active (top), single female speaker active (middle), and both speakers active at the same time. The reverberative patterns are visible in the top and in the middle plots. Furthermore, the patters are seen to overlap in the bottom plot, showing that the reverberation structure for each individual speaker can be recovered (at least visually) when two speakers are active. To verify this, the similarity function described earlier was computed between the direct source and the reflections. It was found that the reflective peaks in the combined image are correctly associated with the corresponding source (i.e., for each reflection signal, the matching score is higher for the corresponding originating source than for the other source).

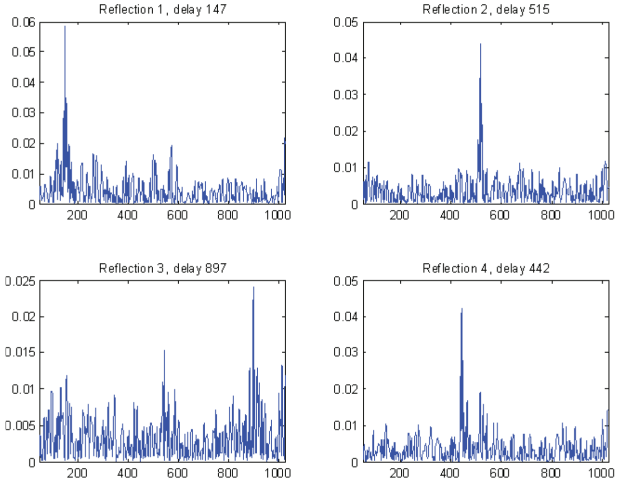

Fig. 4. Evaluation of time delays between the original signal and the reflected versions of it. The values obtained are in very good agreement with true time delays computed using geometric constraints.

After obtaining the similarity metric, the audio image peaks are organized into groups according to the originating source, and the time-delay estimation algorithm is run. The output of the time delay evaluation is shown in Figure 4. To verify the algorithm, the positions of the sources and of the array in the room were measured and the "true" values for time delays were computed using simple geometric reflection models. It was found that the obtained values are each within 4 samples of the time delays estimated by the algorithm. Finally, the beamformed signals are summed up coherently by applying the time delays found.

A plot of the processed signal from the single female speaker recording is shown in Figure 5. In each plot, the data is scaled to full plotting range for details clarity. Addition of the reflections to the original signal increases the signal magnitude; because of the scaling, this is not seen directly but is visible as substantial decrease of the noise level during the beginning silence and in the pauses between individual words, indicating successful dereverberation. The results for the case of two simultaneously active speakers are hard to visualize because the spoken sentences overlap; however, informal listening tests show improvement in speaker separation compared to the single beamformer focused directly at the source. We are currently working on processing larger data set in order to objectively assess SNR improvement and on acquiring data for two speakers who are active in turns to make visualization easier.

6. CONCLUSIONS

We have described a technique for automatic identification of reflections of a sound source off the enclosure walls in order to collect those reflections and to sum them up coherently, similarly to the matched filter array processing proposed before but without the need to actually simulate (or measure) and invert room impulse re-

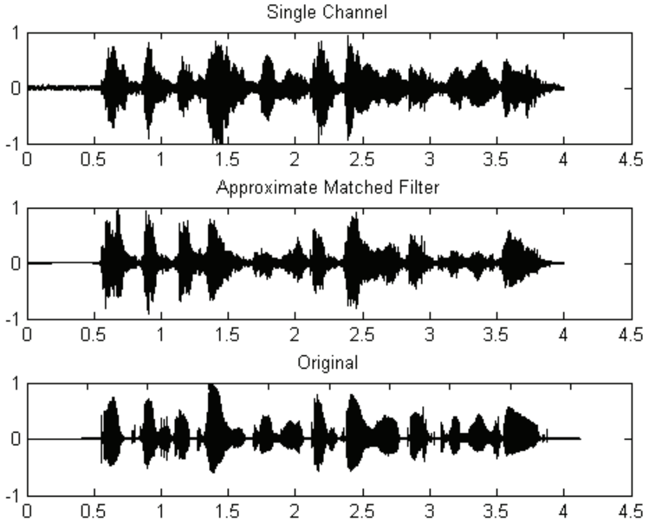

Fig. 5. A single female speaker experiment showing the signal dereverberated using the proposed approximate MFA filter in comparison with the single-channel recording and with the original clean waveform.

sponse. The technique operates by isolating the prominent peaks in the audio image of the room acquired by audio camera and grouping them using the similarity metric (MFCC in this particular case). The time delays between the signals within a group are found, and the signals are summed up coherently. Essentially, the proposed method employs multiple beamformers operating not only on the original source but also on its reflections, enabling one to improve the signal-to-noise ratio in multi-path environments in comparison with traditional beamforming. The experimental evaluation shows that the technique is feasible and that the results obtained are consistent with the expectations. Further evaluations are currently in progress.

7. REFERENCES

[1] J. B. Allen and D. A. Berkeley (1979). "Image method for efficiently simulating small-room acoustics", J. Acoust. Soc. Am., vol. 65, pp. 943-950.

[2] J. Borish (1984). "Extension of the image model to arbitrary polyhedra", J. Acoust. Soc. Am., vol. 75, pp. 1827-1836.

[3] R. Torres, P. Svensson, and M. Kleiner (2001). "Computation of edge diffraction for more accurate room acoustics auralization", J. Acoust. Soc. Am., vol. 109, pp. 600-610.

[4] T. Funkhouser et al. (2004). "A beam tracing method for interactive architectural acoustics", J. Acoust. Soc. Am., vol. 115, pp. 739-756.

[5] S. Sakamoto, A. Ushiyama, and H. Nagatomo (2006). "Numerical analysis of sound propagation in rooms using the finite difference time domain method," J. Acoust. Soc. Am., vol. 120, p. 3008.

[6] S. T. Neely and J. B. Allen (1979). "Invertibility of a room impulse response", J. Acoust. Soc. Am., vol. 66, pp. 165-169.

[7] W. Putnam, D. Rocchesso, and J. O. Smith (1995). "A numerical investigation of the invertibility of room transfer functions", Proc. IEEE WASPAA 1995, New Paltz, NY, pp. 249-252.

[8] B. D. van Veen and K. B. Buckley (1988). "Beamforming: A versatile approach to spatial filtering", IEEE ASSP Magazine, vol. 5, pp. 4-24.

[9] L. J. Griffith and C. W. Jim (1982). "An alternative approach to linearly constrained adaptive beamforming", IEEE Transactions on Antennas and Propagation, vol. AP-30, pp. 27-34.

[10] J. Bitzer, K. Simmer, and K. Kammeyer (1999). "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement", Proc. IEEE ICASSP 1999, Phoenix, AZ, pp. 2965-2968.

[11] E.-E. Jan, P. Svaizer, and J. L. Flanagan (1995). "Matched-filter processing of microphone array for spatial volume selectivity", Proc. IEEE ISCAS 1995, Seattle, WA, pp. 1460-1463.

[12] D. Rabinkin et al. (1998). "Optimal truncation time for matched filter array processing", Proc. IEEE ICASSP 1998, Seattle, WA, vol. 6, pp. 3629-3632.

[13] S. Affes and Y. Grenier (1997). "A signal subspace tracking algorithm for microphone array processing of speech", IEEE Trans. Speech and Audio Proc., vol. 5, pp. 425-437.

[14] S. Gannot and M. Moonen (2003). "Subspace methods for multimicrophone speech dereverberation", EURASIP Journal on Applied Signal Processing, vol. 2003, pp. 1074-1090.

[15] A. E. O'Donovan, R. Duraiswami, and N. A. Gumerov (2007). "Real time capture of audio images and their use with video", Proc. IEEE WASPAA 2007, New Paltz, NY, pp. 10-13.

[16] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein (2001). "Robust localization in reverberant rooms", in *Microphone Arrays: Signal Processing Techniques and Applications*, ed. by M. S. Brandstein and D. B. Ward, Springer-Verlag, Berlin, Germany, pp. 157-180.

[17] J. S. Garofolo et al. (1993). "TIMIT acoustic-phonetic continuous speech corpus", Linguistic Data Consortium, 1993.

[18] J. Meyer and G. Elko (2002). "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield", Proc. IEEE ICASSP 2002, Orlando, FL, vol. 2, pp. 1781-1784.

[19] B. Rafaely (2005). "Analysis and design of spherical microphone arrays", IEEE Trans. Speech and Audio Proc., vol. 13, pp. 135-143.

[20] Z. Li and R. Duraiswami (2007). "Flexible and optimal design of spherical microphone arrays for beamforming", IEEE Trans. Audio, Speech, and Language Proc., vol. 15, pp. 702-714.

[21] A. E. O'Donovan, R. Duraiswami, and J. Neumann (2007). "Microphone arrays as generalized cameras for integrated audio visual processing", Proc. IEEE CVPR 2007, Minneapolis, MN.

[22] A. E. O'Donovan, R. Duraiswami, and D. N. Zotkin (2008). "Imaging concert hall acoustics using visual and audio cameras", Proc. IEEE ICASSP 2008, Las Vegas, NV, pp. 5284-5287.

[23] F. Bimbot et al. (2004). "A tutorial on text-independent speaker verification", EURASIP Journal on Applied Signal Processing, vol. 2004, pp. 430-451.

[24] T. Kinnunen (2005). "Optimizing spectral feature based text-independent speaker recognition", Ph. D. thesis, University of Joensuu, Finland.

[25] Y. Huang, J. Benesty, and G. W. Elko (1999). "Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system", Proc. IEEE ICASSP 1999, Phoenix, AZ, vol. 2, pp. 937-940.