Immersive Audio Synthesis Algorithms

1. Research Team

Project Leader: Prof. Chris Kyriakakis, *Electrical Engineering*

Other Faculty: Prof. Tomlinson Holman, *CNTV*

Graduate Students: C. –S. Lin, A. Mouchtaris, M. Peterson

2. Statement of Project Goals

The work in this project is focused on the synthesis of multichannel immersive audio material. It is connected to the work in the report on rendering algorithms, in that the entire family of synthesis and rendering methods must be considered as a joint optimization problem in order to achieve the desired levels of immersion required by both the 2020Classroom and the Distributed Performance immersipresence projects.

3. Project Role in Support of IMSC Strategic Plan

Immersive audio is one of the critical elements in IMSC's vision of Immersipresence and is intricately tied in to several projects at many levels. These include the 2020Classroom project and its Media Immersion Distance Learning Classroom (MIDLeC), the Distributed Immersive Performance project, as well as the Remote Media Immersion implementation within the MIE architecture.

4. Discussion of Methodology Used

Virtual Microphone Signal Synthesis Algorithms

The need for low-latency two-way multichannel audio communication for both the 2020Classroom and Distributed Performance applications imposes strict requirements. Traditional methods of compression (*e.g.,* Dolby AC-3, MPEG, and AAC) involve computation times during the encoding process that are on the order of 200-500 μs, thus making them unsuitable for real-time communication. An approach that we have proposed and are continuing to expand on is to synthesize the multiple channels of audio needed for realistic immersion at the receiving end from a smaller number of signals (*e.g.* the left and right stereo signals).

While this approach has proven very successful in synthesizing signals that would be present in distant microphones (capturing the reverberant sound in the space), the synthesis of spot microphone signals that would be placed close to a musical instrument or person remains a difficult problem. It should be noted that this is not the same as the source separation type of problem in which it is desired to suppress all other sound except for one particular instrument or voice. Instead, it is a problem of creating a signal exactly as it would be recorded by a closely-place microphone–it contains mostly sound from the source of interest, but also other sound depending from other nearby sources.

The approach we followed is that of spectral conversion. A training data set is created from reference and target recordings in the space of interest by applying a short sliding window and extracting the parameters that model the short-term spectral envelope (*e.g.*, cepstral coefficients). This set is created based on the parts of the target recording that we desire to enhance in the final recording. If, for example, we desire to enhance speech then the training set is created using speech elements. The training procedure produces a reference and a target vector sequence and the goal is to find a function that we can apply to the reference sequence and produce the target sequence. The method we used to find this function is based on Gaussian mixture models (GMM) that allow us to model each sequence as a sum of random vectors with probability densities that represent a weighted sum of normal multivariate distributions. A cost function is then minimized in the least squares sense to find the appropriate weights for the signals of interest.

In our experiments we used a data set of 10,000 spectral vectors and then applied the method described above to generate a synthesized version of the desired target signal. The synthesized version was then compared to the actual signal recorded by a real microphone in the location we were attempting to synthesize. A series of listening tests (A/B/X) and objective performance criteria were used to evaluate the performance. Our results indicate that the GMM model works well, particularly when the signal is divided into sub bands. For the case of speech, we emphasized the 100 Hz – 5 kHz regions and also varied the order of the LPC filter in each band so that we could better approximate the variations in the signal.

5. Short Description of Achievements in Previous Years

- First group to propose the use of virtual microphone algorithms
- Filters designed for various representative spaces that could be used to convert mono and stereo recordings to multichannel versions
- Real time synthesis of distant microphones demonstrated
- Off-line synthesis of spot microphones for percussive sounds demonstrated using Choi-Williams distributions

5a. Detail of Accomplishments During the Past Year

- Gaussian mixture model proposed for general-purpose spot microphone synthesis
- Experimental verification of GMM method for speech enhancement applications
- Code optimization for porting to 2020Classroom two-way communication scenario

6. Other Relevant Work Being Conducted and How this Project is Different

Related work in this area falls into two categories. The first has to do with attempts to create multichannel recordings from mono or stereo material. Here, all of the methods that exist both in industry and in the literature are based on various forms of artificial reverberation. Whether these use sophisticated room modeling programs (e.g., CATT acoustics) or elaborate banks of all-pass filters, they are fundamentally different from our approach in that they "add" reverberation to signals that already have reverberation in them from the original recording. This

produces audible artifacts. Our method is similar to morphing of images in that we transform the signal from one microphone into that of another by incorporating the relevant acoustical and spectral characteristics. Although the existing model-based methods allow the user to easily change the modeled environment (*e.g.,* from a small room to a large hall) they still are not able to overcome the tinny sound that is characteristic of the comb-filters used to produce the reverberant sound. For the spot microphone case that we have been working, there is no other solution or related work that we are aware of. The source separation methods that are being investigated deal with the problem of isolating one source from noise or from other sources, but in no way attempt to synthesize the signal in a closely-placed microphone as would be done in a multichannel recording.

7. Plan for the Next Year

Transient sounds still present a significant challenge because they cannot be modified by simply processing the short-term spectral envelopes. We plan to focus on such sounds next so that we can add to our library of algorithms for multichannel audio synthesis. We also plan to demonstrate a real-time streaming system that uses today's two-channel streaming infrastructure and creates a multichannel version of the content at the receiving end in real time.

8. Expected Milestones and Deliverables

A stand-alone, DSP-based solution for processing two-channel audio into multichannel will be developed by porting our algorithms to the TI DSP platform. This will allow us to off-load the computation from the client computer and thus increase the length of the filters for achieving better synthesis performance.

9. Member Company Benefits

N/A

10. References

[1] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Trans. Speech and Audio Processing*, 3(5): 357–366, September 1995.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 655–658, New York, NY, April 1988.

[3] A. Mouchtaris and C. Kyriakakis. Time-frequency methods for virtual microphone signal synthesis. In 111*th Convention of the Audio Engineering Society (AES), preprint No. 5416*, New York, NY, November 2001.

[4] D. A. Reynolds and R. C. Rose. Robust text independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, 3(1): 72–83, January 1995.