Final Program and Paper Summaries, Workshop Applications Signal Processing, Audio, Acoustics (New Paltz, NY), pp. 139-140, Oct. 1991

Efficient Methods for Simulating a Moving Talker in a Rectangular Room

Benoit Champagne, Arthur Lobo and Peter Kabal

INRS-Télécommunications, Université du Québec, 3 Place du Commerce
Verdun, Québec, Canada H3E 1H6 (514 765-7844)

Abstract

In this paper, we describe two methods for efficiently simulating the response of a microphone to a moving talker in a rectangular room. Both methods are based on an extension of the image method to moving sources. In the first method, the microphone output signal is obtained by performing a time-domain filtering operation on the original speech signal, while in the second method, a time-frequency representation of this filtering operation is used. In each case, computational load and memory requirements are considerably reduced by taking advantage of the fact that the talker velocity is much smaller than the speed of sound.

I. Introduction

In applications such as teleconferencing where it is necessary to monitor a speech signal in the presence of reverberation and interfering noise sources, the use of a microphone array can lead to significant improvements in the intelligibility of the captured speech signal, when compared to the use of a conventional single-microphone monitoring device [1]. This is due mainly to the high directionality of the array which attenuates acoustic waves coming in from unwanted directions. Such directionality is achieved by the application of signal processing techniques which, in the simplest form, consist of introducing time delays in the microphone outputs prior to adding them coherently.

A common requirement of practical microphone array systems is the ability to locate and track a talker as he/she moves through a room. In this respect, computer generation of synthetic microphone output signals corresponding to those generated by a moving talker in a room is of considerable interest. Indeed, it enables the system designer to test, adjust the parameters of, and eventually select particular array processing algorithms, without the actual need (and cost) of setting up an experimental facility.

In this paper, we present two methods for efficiently simulating a moving talker in a rectangular room. Given the speech signal, the trajectory of the talker, the characteristics of the room and the microphone location, both methods produce an accurate approximation to the signal that would be picked up by the microphone under perfect wave propagation conditions. Both methods are based on an extension of the image method [2] to moving sources. In the first method, the microphone output signal is obtained by performing a time-domain filtering operation on the original speech signal, while in the second method, a time-frequency representation of this filtering operation is used. In each case, computational load and memory requirements are considerably reduced by taking advantage of the fact that the talker velocity is much smaller than the speed of sound.

II. Time-domain representation

The image method, used in [2] to simulate the microphone output generated by a fixed talker, can easily be extended to the moving talker case. Denoting by $a(t)$ the original speech signal at time t and by $s(t)$ the corresponding microphone output signal (properly scaled), we obtain

$$s(t) = \sum_r \frac{\beta_r}{\tau_r(t)} a(t - \tau_r(t)) \qquad (1)$$

where the summation is over all image indices r, β_r is the attenuation coefficient for image r, and $\tau_r(t)$ is the non-integer propagation delay for an acoustic wave originating from image r and reaching the microphone at time t. (In practice, $a(t)$ is a discrete time signal and $a(t-\tau_r(t))$ is obtained by band-limited interpolation [3].) This propagation delay is defined implicitly by the equation

$$c\tau_r(t) = g_r(t - \tau_r(t)) \qquad (2)$$

where c is the speed of sound in air and $g_r(t)$ is the distance between the microphone and the image r at time t.

The major difficulty with the time-domain filtering operation (1) on $a(t)$ is posed by $\tau_r(t)$. Indeed, because of its time-varying nature, the propagation delay must be precomputed for each value of time t and each image r considered. Moreover, each of these computations requires the solution of the implicit equation (2) by an iterative algorithm. The "exact" evaluation of $s(t)$ by means of (1), together with an appropriate iterative algorithm for (2), is computationally very expensive.

However, the above considerations do not take into account the fact that the talker velocity is generally much smaller than the speed of sound in air. If the talker velocity is sufficiently small, it can be shown that the solution of (2) is given explicitly, and with very good accuracy, by

$$c\tau_r(t) = g_r(t - \mu_r) \tag{3}$$

where $\mu_r = g_r(0)/c$ represents the travel time between the image r and the microphone, for a wave originating from image r at time 0. Moreover, because of the slowly-varying nature of the functions $g_r(t)$, it turns out that it is not necessary to solve (3) for all values of t. Hence, it is possible to save on memory requirement by evaluating $\tau_r(t)$ on a widely spaced grid of values of t and by using an on-line interpolation routine to evaluate the missing values of $\tau_r(t)$. For this purpose, we have found that linear interpolation is sufficient.

The above approach results in large savings in computation time and memory requirement. It can be verified that this approach remains valid as long as the following condition is satisfied:

$$BLv \ll c^2 \tag{4}$$

where B is the highest frequency (in Hertz) of the speech signal, L is the room main diagonal, and v is the maximum talker velocity. Since the rate of change of $\tau_r(t)$ is bounded by v/c, (4) can be interpreted as a requirement that the change in the phase term $B\tau_r(t)$, during the time required by a wave to propagate through the room, L/c, remain much smaller than 1 for all images r.

III. Time-frequency domain representation

Equation (1) can also be used to express the microphone output signal $s(t)$ as the convolution of $a(t)$ with an appropriate impulse response $h(t, u)$. Although $h(t, u)$ is generally not time-invariant, it can be written in the form $f(t - u; u)$ where, in the case of a moving talker, $f(.,.)$ is slowly varying in its second argument. This suggests the use of the following time-frequency representation [4] of $h(t, u)$,

$$H(t, \omega) = \int h(t, u)e^{j\omega(t-u)}du \tag{5}$$

which can be interpreted as a time-varying frequency response.

Let $A(\omega)$ denote the Fourier transform of the original speech signal $a(t)$ and $H_n(\omega)$ denote the Fourier coefficients of $H(t, \omega)$ (for a fixed frequency), i.e.,

$$H_n(\omega) = \frac{1}{T}\int_0^T H(t, \omega)e^{-j2\pi t/T}dt \tag{6}$$

where T is the observation interval. An exact expression for $s(t)$ in terms of $H_n(\omega)$ and $A(\omega)$ can be obtained as follows

$$s(t) = \frac{1}{2\pi}\sum_n \{\int H_n(\omega)A(\omega)e^{j\omega t}d\omega\}e^{j2\pi nt/T} \tag{7}$$

The bracketed quantity represents the output of a time invariant linear system with impulse response $H_n(\omega)$. In general, the evaluation of $s(t)$ through (7) is not very practical since the summation over n extends from $-\infty$ to ∞ and since the calculation of each of the $H_n(\omega)$ is computationally expensive.

In the case of a slowly moving talker (i.e. $BLv \ll c^2$), however, the approximations used in Section II can also be used to simplify considerably the evaluation of the Fourier coefficients $H_n(\omega)$. But, more importantly, it can be verified that $H_n(\omega) \approx 0$ for

$$n > \frac{v}{c}BT \tag{8}$$

Hence, only a relatively small number of Fourier coefficients $H_n(\omega)$ need to be evaluated. The use of (7) for the evaluation of $s(t)$ therefore becomes very efficient in the case of a moving talker.

IV. Discussion

In a preliminary experiment on tracking algorithms for microphone arrays, the time-domain representation method was used to generate the output signals of two spatially separated microphones monitoring a talker moving on a straight line at constant velocity. The microphone outputs were then used to test an adaptive time delay tracker based on the LMS algorithm. The flexibility of the synthetic signal generation method enabled us to test the tracker for different talker velocities, microphone positions, and wall reflection coefficients.

A comparative study, in terms of computational and memory requirements, of the two methods presented in this paper is currently in progress.

Acknowledgement

This work was supported in part by Bell Northern Research, Ottawa, Canada.

References

[1] J. L. Flanagan, J. D. Johnston, R. Zahn, G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," J. Acoust. Soc. Am., vol. 78, pp. 1508–1518, Nov. 1985.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am., vol. 65, pp. 943–950, Apr. 1978.

[3] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," J. Acoust. Soc. Am., vol. 80, pp. 1527–1529, Nov. 1986.

[4] L. A. Zadeh, "Frequency analysis of variable networks," Proc. I. R. E., vol. 38, pp. 291–299, March 1950.