ACOUSTIC BEAMFORMING EXPLOITING DIRECTIONALITY OF HUMAN SPEECH SOURCES

Terence Betlehem and Robert C. Williamson

RSISE, Australian National University
Canberra, ACT, 0200, Australia
[Terence.Betlehem, Bob.Williamson]@anu.edu.au

ABSTRACT

This paper examines the improvement that can be attained with perfect knowledge of the sound source directivity pattern and orientation in beamformer designs in the problem of speech acquisition. Data-independent beamformers are derived through formulation of a constrained optimization problem with a unity-gain constraint. Using computer simulation, these beamforming schemes are compared to the delay and sum (DS) beamformer and the best single sensor in a reverberant room environment. Criteria used to measure performance are (1) the direct to reverberant ratio, to assess extent of reverberation suppression, and (2) an objective measure of speech intelligibility called the speech transmission index (STI). For human-speaker source directivity, simulation results show that modest improvements to performance are obtainable.

1. INTRODUCTION

Over the last 30 years, much work has been done in the area of array signal processing. Unfortunately, many of the results developed have a number of constraints. Conventional theory assumes sensors are evenly arranged, sources lie in the far-field and signals are narrowband. Such restrictions are inadequate for many speech acquisition problems. To remedy such shortcomings, theory has recently been developed for broadband sources in the near-field [1]. Other work presents a framework for beamforming to a source located amongst an array of randomly placed sensors [2].

In this paper we further extend the generality of beamformer designs by accounting for the directivity of the source. This consideration is irrelevant for clustered arrays in the far-field as each sensor experiences the same directivity factor. However for sparse arrays, sensors could be, for example, both in front of and behind the source. For human speakers, sensors behind the source can experience over 15dB in signal attenuation at high frequencies [3]. In such cases the beamforming scheme should apply more weight to sensors in front of the source.

We present two beamformer designs. These designs take advantage of perfect knowledge of the source directivity pattern and orientation. They also approximately maximize the speech intelligibility criterion proposed by Thiele [4]. We then compare their performance with other beamforming schemes under computer simulation of reverberant room conditions and the human-speaker source directivity pattern [3]. We do this with both a measure of reverberation suppression and an objective measure of speech intelligibility called the speech transmission index [5].

2. DIRECTIONAL-SOURCE BEAMFORMING

Consider a directional sound source $S(\omega)$ placed in a reverberant room amongst an array of N omnidirectional microphones. The microphones are located at arbitrary points $\boldsymbol{x}_n$, $n = 1, ..., N$. The source is at point $\boldsymbol{s}$. Let $Q_n(\omega)$ denote the directivity factor of the source in the direction of each microphone as a function of frequency ω.

Beamforming involves filtering the output of each microphone by a filter $H_n(\omega)$ and summing the result. To perform beamforming, we attach to the microphones the filters $H_n(\omega)$. Ignoring sensor self-noise and interfering noise sources, the output of the beamformer can be written as

$$Y(\omega) = \sum_{n=1}^{N} G_n(\omega) H_n(\omega) S(\omega),$$

where $G_n(\omega)$ is the transfer function between the source and sensor n. The source to beamformer transfer function $Y(\omega)/S(\omega)$ can be written as $\boldsymbol{g}^H \boldsymbol{h}$ where $\boldsymbol{g}^H \triangleq [G_1(\omega), G_2(\omega), ..., G_N(\omega)]$, $\boldsymbol{h}^T \triangleq [H_1(\omega), H_2(\omega), ..., H_N(\omega)]$, and $(\cdot)^H$ is the complex conjugate transpose operator. For compactness, dependence on ω will be suppressed for the rest of the paper.

We express the room transfer function as the sum of the direct part $\boldsymbol{g}_d$ and the reverberant part $\boldsymbol{g}_r$. Neglecting room diffraction effects, the elements of $\boldsymbol{g}_d$ are equal to the free field transfer function for an omnidirectional source scaled by the source directivity factor Q_n:

$$[\boldsymbol{g}_d]_n = \frac{Q_n}{d_n} e^{-j\frac{\omega}{c}d_n}, \quad n = 1, \ldots, N \tag{1}$$

where $[\boldsymbol{g}_d]_n$ is the nth element of $\boldsymbol{g}_d$, $d_n \triangleq ||\boldsymbol{s} - \boldsymbol{x}_n||$ is the distance from the source to sensor n, and c is the speed of sound.

The reverberant part $\boldsymbol{g}_r$ is much more difficult to model. The fine structure of the reverberant field is strongly dependent on the geometry and material of the room boundaries. However, two basic models of room reverberation are available, namely the diffuse sound field [4] and the image-source method.

Conventional beamformer design focusses on the direct part. One can successfully beamform the direct part without prior knowledge of the parameters of the room. Consequently, it is common to constrain the direct part of the beamformer to reproduce the original input signal with no distortion. This is equivalent to writing

$$\boldsymbol{g}_d^H \boldsymbol{h} = 1. \tag{2}$$

3. BEAMFORMER DESIGN

Ideally we consider the beamformer design problem of choosing $\boldsymbol{h}$ in order to maximize speech intelligibility subject to (2). This requires an objective measure of intelligibility.

One measure due to Thiele [4] is based on the observation that reflections with a delay time less than 50 ms improve speech intelligibility. By spatially selecting sound less than 50ms, or 17 m, away from the sensors (using $c = 342$ m/s), we can maximize Thiele's intelligibility criterion. Section 3.1 presents a design that approximately does this. Section 3.2 presents a design criteria for minimizing uncorrelated sensor noise. Both designs can be shown to be identical in the case of well-separated sensors (i.e. $\omega d_{mn}/c \gg 1$).

3.1. Minimum Far-field Power

A beamformer design that approximately maximizes Thiele's criterion is presented in [2]. Here, we minimize the output power of the beamformer to far-field isotropic noise subject to (2). This reduces to [2]:

$$\min_{\boldsymbol{h}} \boldsymbol{h}^H \boldsymbol{R} \boldsymbol{h} \quad \text{subject to} \quad \boldsymbol{g}_d^H \boldsymbol{h} = 1,$$

where the sensor correlation matrix $\boldsymbol{R}$ is given by

$$[\boldsymbol{R}]_{m,n} = \text{sinc}\left(\omega d_{mn}/c\right), \quad m, n = 1, \ldots, N, \tag{3}$$

$\text{sinc}(x) \triangleq \sin x/\, x$, and $d_{mn} \triangleq ||\boldsymbol{x}_m - \boldsymbol{x}_n||$ is the sensor-sensor spacing. This constrained linear optimization problem has the well known solution

$$\boldsymbol{h}_{ff} = \frac{\boldsymbol{R}^{-1}\boldsymbol{g}_d}{\boldsymbol{g}_d^H \boldsymbol{R}^{-1} \boldsymbol{g}_d}. \tag{4}$$

3.2. Minimum White Noise Gain

For comparison, we derive another (simpler) beamformer which minimizes the white noise gain (WNG) $\boldsymbol{h}^H \boldsymbol{h}$ under the same constraint. WNG is the output power due to unit variance white noise at the sensors. This is equivalent to setting $\boldsymbol{R} = \boldsymbol{I}$, and thereby assuming interfering noise is uncorrelated between sensors. The solution is simply

$$\boldsymbol{h}_{wng} = \frac{1}{\boldsymbol{g}_d^H \boldsymbol{g}_d} \boldsymbol{g}_d. \tag{5}$$

In terms of original parameters

$$[\boldsymbol{h}_{wng}]_n = \frac{1}{\boldsymbol{g}_d^H \boldsymbol{g}_d} \frac{Q_n}{d_n} e^{-j\frac{\omega}{c} d_n}.$$

Contrast this with the delay and sum (DS) beamformer

$$[\boldsymbol{h}_{ds}]_n = e^{-j\frac{\omega}{c} d_n}. \tag{6}$$

Ignoring the frequency dependent normalization term $(\boldsymbol{g}_d^H \boldsymbol{g}_d)^{-1}$ (recall we suppress dependence on ω to simplify notation), the WNG beamformer is similar to the DS beamformer in that it time aligns the direct parts of the sensor signals. However, the WNG beamformer applies more weight to the sensors that (a) experience greater source directivity and (b) are closer to the source.

4. PERFORMANCE MEASURES FOR BEAMFORMERS

We now describe the measures of performance of the beamformers. Section 4.1 defines the direct-to-reverberant ratio (DRR) that is used to quantify reverberation suppression. Section 4.2 summarizes speech transmission index (STI) that is used to quantify speech intelligibility. Both DRR and STI can be applied directly to the room/beamformer impulse response.

4.1. Direct-to-Reverberant Ratio

The DRR of an impulse response is the ratio of the direct part energy to the reverberant part energy. For the output of a beamformer, DRR is given by

$$\gamma = \frac{\int_{-\infty}^{\infty} |S(\omega)|^2 |\boldsymbol{g}_d^H(\omega)\boldsymbol{h}(\omega)|^2 \, \mathrm{d}\omega}{\int_{-\infty}^{\infty} |S(\omega)|^2 |\boldsymbol{g}_r^H(\omega)\boldsymbol{h}(\omega)|^2 \, \mathrm{d}\omega}. \tag{7}$$

$|S(\omega)|^2$ has been set to 1 in the frequency range 100Hz-10kHz, zero otherwise. DRR is a reasonable measure of reverberation suppression for beamformers. However it is not able to quantify intelligibility-related effects such as syllabic blurring.

4.2. Speech Transmission Index

The speech transmission index (STI) is an objective measure of speech intelligibility over acoustic channels [5]. STI determination can be summarized into six steps:

1) Modulation Transfer Function: The MTF is defined as the modulation index of the intensity envelope of a transmitted test signal. The modulation index measures the blurring of syllables occurring over the reverberant channel that reduces speech intelligibility. The test signal comprises octave-bandlimited white noise amplitude modulated with the function $\sqrt{1 + \cos(2\pi F t)}$, where F is the modulation frequency.

Over purely reverberant channels, the MTF can be calculated analytically for a test signal with unfiltered white noise from the Fourier transform of the room impulse response squared [6]:

$$m(F) = \frac{\int_0^{\infty} h^2(t) e^{2\pi F t} \, \mathrm{d}t}{\int_0^{\infty} h^2(t) \, \mathrm{d}t}. \tag{8}$$

For the MTF used in STI, the white noise is octave-bandlimited by filtering with 6th order Butterworth filters of center frequencies from 125Hz to 8kHz (IEC standard 60268-16). For octave-bandlimited noise, it is easy to show that (8) is still valid, provided that (a) $h(t)$ is obtained by convolving the room or beamformer impulse response with the impulse response of the bandpass filter; and (b) the lower bandpass cutoff frequency is much larger than maximum modulating frequency.

2) Conversion to Effective SNRs: The STI definition requires $m(F)$ for each octave be sampled for 14 modulation frequencies F_n in the range 0.5 to 16 Hz, spaced in $1/3$-octave intervals. Each modulation index is then converted to an effective SNR through the transformation,

$$\text{SNR}_i(F_n) = 10 \log \left(\frac{m_i(F_n)}{1 - m_i(F_n)} \right),$$

where $i = 1 \ldots 7$ and $n = 1 \ldots 14$. Subscript i refers to the octave band.

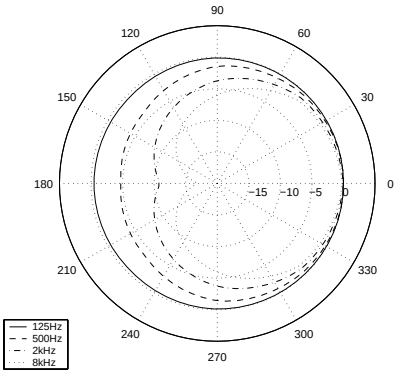

Fig. 1. Human speaker directional response (dB) obtained by least squares fitting to [9]. Average directivity index (100Hz - 10kHz) is 4.4 dB.

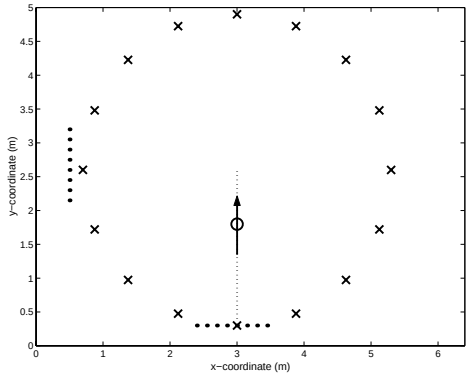

Fig. 2. Location of source (∘), circular array sensors (×) and linear array sensors (·) for simulations. DRR and STI were measured along the dotted line. Source faced north.

3) Range Limiting: There are lower and upper SNR limits outside of which negligible difference is made to speech intelligibility. Effective SNR is hence hard limited to the range ± 15 dB.

4) Octave-Band-Specific SNRs: To combine the 14 effective SNRs of each octave, they are simply averaged:

$$\overline{\mathrm{SNR}}_i = \frac{1}{14}\sum_{n=1}^{14} \mathrm{SNR}_i(F_n).$$

5) Transmission Indices: Transmission indices are obtained by scaling effective SNRs of each octave to the range 0 to 1:

$$\mathrm{TI}_i = \frac{\overline{\mathrm{SNR}}_i + 15}{30}.$$

6) Octave Weighting: Finally, the STI is obtained by applying weighting factors α_i to the octave-band-specific SNRs and summing. Recent findings [7] have found that contributions from different frequency bands are not purely additive, and suggest the inclusion of redundancy correction factors β_i:

$$\begin{aligned}\mathrm{STI} = \alpha_1\mathrm{TI}_1 - \beta_1\sqrt{\mathrm{TI}_1\mathrm{TI}_2} + \alpha_2\mathrm{TI}_2 \\ - \beta_2\sqrt{\mathrm{TI}_2\mathrm{TI}_3} + \ldots + \alpha_7\mathrm{TI}_7\end{aligned}$$

The values of α_i and β_i used were obtained from the female speaker results of [7], where the STI was fit to CVC-word score data.

5. DIFFUSE FIELD DIRECT TO REVERBERANT RATIO

We now apply (7) to calculate the DRR of a beamformer in the case where reverberation is modelled by a diffuse field. In a diffuse field, reverberation term $\boldsymbol{g}$ becomes a random process. Consequently, we replace the matrix $\boldsymbol{g}_r\boldsymbol{g}_r^H$ in (7) with its expectation. Now, the sensor-sensor spatial correlation given by (3). Also, energy density $E_r(\omega) \triangleq \mathrm{E}\{[\boldsymbol{g}_r]_m[\boldsymbol{g}_r]_n\}$ is independent of sensor location. It can be hence shown that $\mathrm{E}\{\boldsymbol{g}_r\boldsymbol{g}_r^H\} = E_r\boldsymbol{R}$. The diffuse field DRR is hence

$$\gamma = \frac{\int_{-\infty}^{\infty} |S(\omega)|^2 |\boldsymbol{g}_d^H(\omega)\boldsymbol{h}(\omega)|^2 \,\mathrm{d}\omega}{\int_{-\infty}^{\infty} |S(\omega)|^2 E_r(\omega)\boldsymbol{h}^H(\omega)\boldsymbol{R}(\omega)\boldsymbol{h}(\omega)\,\mathrm{d}\omega}. \qquad (9)$$

For the DS beamformer, the direct part $\boldsymbol{g}_d^H\boldsymbol{h}$ is distorted by factor $\sum_{n=1}^{N}\frac{Q_n}{d_n}$. If the array is sparse, i.e. sensor-sensor spacings are "large" over the frequency range of interest ($\omega d_{mn}/c \gg 1$), we can set $\mathrm{sinc}(\omega d_{mn}/c) \approx \delta_{mn}$ so that $\boldsymbol{R} \approx \boldsymbol{I}$. Under this assumption, $\boldsymbol{h}^H\boldsymbol{R}\boldsymbol{h}$ reduces to $1/\boldsymbol{g}_d^H\boldsymbol{g}_d$ for FF and WNG beamformers. For a sparse array, the FF case is equivalent to the WNG case.

From geometric room acoustics [4] diffuse field energy density $E_r(\omega) = \frac{\overline{Q^2}}{16\pi A}\frac{1-\overline{\alpha}}{\overline{\alpha}}$ where $\overline{Q^2}(\omega)$ is the square directivity factor averaged over all directions, $\overline{\alpha}$ is average wall absorption coefficient and A is wall surface area.

6. SIMULATION

We simulated the above beamformers operating in a rectangular room of dimensions $6.4 \times 5 \times 4$m with wall absorption coefficient $\alpha = 0.2$, $c = 342$ m/s, yielding a reverberation time of 0.7 s. In each simulation, we compared the DRR and STI of the DS beamformer, the minimum WNG beamformer with perfect knowledge of the directional source (labelled DWGN), the minimum WNG beamformer working on the assumption that the source is omnidirectional ($Q_n = 1$; labelled OWGN) and the best sensor reading. For further comparison, the diffuse field DRR expression of (9) in each case was also plotted.

To emulate the directivity of a human speaker, the source directional response used was obtained by least squares fitting to Dunn and Farnsworth 60cm xy-plane sound field pressure data [3] (Fig. 1). The room impulse response was determined with the image-source method, assuming a directional source and omnidirectional images. Room impulse responses were bandlimited to the 100Hz - 10kHz frequency range. The source and the sensors were placed 2.5m above the floor of the room.

6.1. Circular Array

Simulation was performed on a 16-element circular array of radius 2.2m centered at $(3, 2.6, 2.5)$ (Fig. 2). In Fig. 3(a), "best mic." and OWNG are seen to outperform DWNG at small source-sensor distance ($d < 0.2$m). Because of the unity-gain constraint, DWNG compensates for the low frequency rolloff caused by positioning a sensor behind the source, by amplifying high frequencies at the expense of reduced DRR improvement. However for larger distances, DWNG outperforms all other beamforming schemes–by up to 1.8dB.

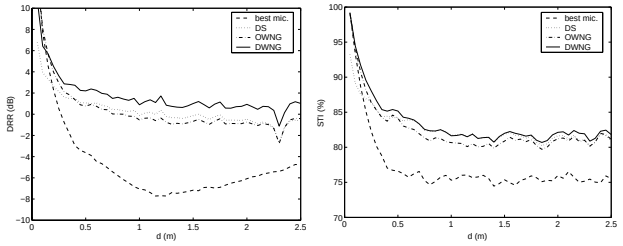

Fig. 3. Output DRR and STI for various beamforming schemes using the circular array.

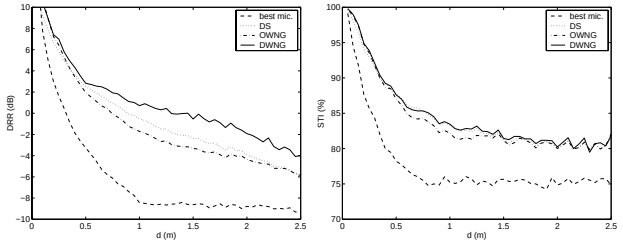

Fig. 4. Output DRR and STI for various beamforming schemes using the linear array.

In contrast, DWNG yields best speech intelligibility at every distance. This shows there is more to improving speech intelligibility than simply maximizing DRR–additional improvement in this case was obtained by imposing the unity-gain constraint. DWNG yields up to a 2% STI improvement over other methods at large distances. Such an improvement is significant, as most STI improvements of using beamformers over the "best mic." are only of order 7% anyway.

6.2. Pair of Linear Arrays

Simulation was performed on a pair of 8-element linear arrays with 0.15m sensor-sensor spacings (figure 4). We see here that DWNG significantly outperforms all other schemes in DRR. DS and OWNG performance tended to stay below DWNG at larger distances, as these designs apply excessive weighting to the sensors in the linear array that the source is facing away from. (Recall DS weights all sensors equally.)

In STI, DWNG is not as impressive - it performs no better than DS. It does outperform the OWNG beamformer by up to 1% at some distances, however.

6.3. Diffuse Field

For further comparison, the diffuse field DRR expression of (9) has been plotted for both array geometries (Fig. 5). These plots preserve the general curve shapes and trends of Fig. 3 and Fig. 4.

7. CONCLUSION

Source directivity has been included into beamformer designs in a straight-forward manner. The minimum WNG beamformer de-

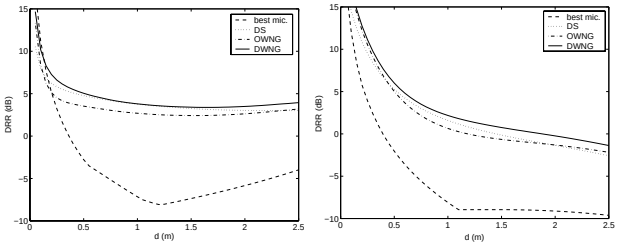

Fig. 5. Diffuse field DRRs for circular and linear array cases.

rived in this paper was shown to provide a "best of both worlds" solution. It was shown to outperform the DS beamformer at small source-sensor distances, and outperform the "best sensor" criterion at large source-sensor distances. Furthermore, inclusion of the source directivity has been shown to improve beamformer performance. In the circular array examples presented, we obtained up to an additional 1.8dB reverberation suppression and 2% STI improvement over the other beamforming techniques when beamforming to a human-speech source.

This work has highlighted the need to use an objective measure of speech intelligibility to assess beamformer performance. Simulation results have shown that, what may lead to significant improvement to direct-to-reverberant ratio does not always improve in speech intelligibility to the same extent.[1]

8. REFERENCES

[1] T.D. Abhayapala, R.A. Kennedy, and R.C. Williamson, "Nearfield broadband array design using a radially invariant modal expansion", *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 392–403, 2000.

[2] D.B. Ward and R.C. Williamson, "Beamforming for a source located in the interior of a sensor array", Proceedings of the Fifth International Sympoisium on Signal Processing and its Applications, vol. 2, pp. 873–876, 1999.

[3] H.K. Dunn and D.W. Farnsworth, "Exploration of pressure field around the human head during speech", *J. Acoust. Soc. Amer.*, vol. 10, no. 1, pp. 184–199, 1939.

[4] H. Kuttruff, *Room Acoustics*, Applied Science Publishers, London, 2nd edition, 1979.

[5] T. Houtgast and H.J.M. Steeneken, "Review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria", *J. Acoust. Soc. Amer.*, vol. 77, no. 3, pp. 1069–1077, 1985.

[6] T. Houtgast, H.J.M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. I. general room acoustics", *Acustica*, vol. 46, pp. 60–72, 1980.

[7] H.J.M. Steeneken and T. Houtgast, "Mutual dependence of the octave-band weights in predicting speech intelligibility", *Speech Communication*, vol. 28, pp. 109–123, 1999.

[1]This work was supported by the Australian Research Council.