# Chapter 6

# Acoustic Beamforming Exploiting Directionality of Human Speech Sources

## 6.1 Introduction

Over the last 30 years, much work has been done in the area of beamforming. Theory has been developed for arrays of irregular geometry, nearfield sources, and for broadband signals. In this chapter, we explore an additional consideration to beamforming - the directionality of the sound source.

Although elaborate theory has recently been developed in the area of nearfield beamforming [1, 2], several illuminating beamformers have been designed through least squares style optimization techniques. Using least squares methods, Ryan and Goubran [80] demonstrated the merits of nearfield designs, showing that beamformers can provide a degree of range discrimination. Ward and Williamson [109] designed a beamformer that exploited range discrimination to minimize the amount of harmful reverberation to the intelligibility of captured speech signals. By using a constrained optimization approach, their design allowed arbitrary placement of sensors.

An important measure of the performance of acoustic systems is the speech intelligibility. Speech intelligibility has traditionally been quantified through subjective listening tests, but more recently accurate objective criteria have been devised. Noting that the late echoes of reverberation are detrimental to speech intelligibility while early echoes tend to reinforce the speech signal, Thiele [101] proposed the fraction of early energy. Steeneken and Houtgast [42] extended the well-known Articulation Index of French and Steinberg [31] to handle reverberation. Called the speech transmission index (STI), it is based on measurement of the modulation transfer function (MTF) [41] of an acoustic channel. Recently they presented a revised method for the STI [90] that is closely correlated with results of consonant-

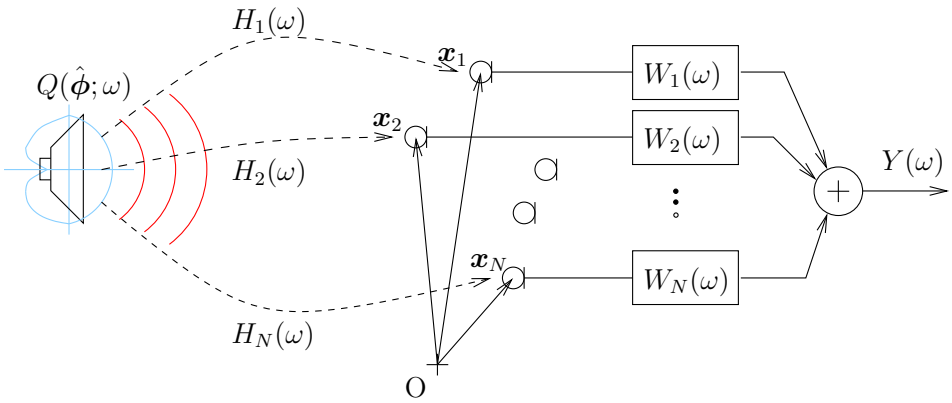



Figure 6.1: Configuration of a directional source with radiation pattern $Q(\hat{\boldsymbol{\phi}};\omega)$ and sensor array for beamformer. Each sensor $n$ is positioned at $\boldsymbol{x}_n$ and attached to a filter with the frequency dependent weight $W_n(\omega)$. Acoustic transfer functions $H_n(\omega)$ are also shown.

vowel-consonant (CVC) word score tests [92].

In this chapter, we examine the improvements made possible by designing beamformers that exploit perfect knowledge of the radiation pattern and orientation of the source. We assess these improvements for the interior field beamformer of Ward and Williamson [109]. Specifically considered in this chapter is a human speaker radiation pattern. Human speaker directionality is quite significant at higher frequencies, increasing up to 15dB at 8kHz [24]. Though not relevant to farfield beamformer design, directionality of the source becomes relevant when microphone arrays are in the nearfield. Tests are performed under simulation of reverberant room conditions and performance assessed with the STI.

Previewing the chapter content, Section 6.2 summarizes the room-beamformer model. Section 6.3 describes the beamformer designs. Section 6.4 details the measures used to assess performance. Section 6.6 assesses the performance of the beamformer designs, using key examples.

## 6.2 Directional-Source Beamforming

Consider a directional sound source placed in a reverberant room amongst an array of $N$ omnidirectional microphones, located at arbitrary points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ as depicted in Figure 6.1. The source is located at point $\boldsymbol{y}$ and possesses a radiation pattern $Q(\hat{\boldsymbol{\phi}};\omega)$ in each direction $\hat{\boldsymbol{\phi}}$.

Beamforming involves filtering the output of each microphone by a filter $W_n(\omega)$ and summing the result. Ignoring sensor self-noise and interfering noise sources,

the output of the beamformer can be written as

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

where $H_n(\omega)$ is the acoustic transfer function between the source and sensor $n$ and $S(\omega)$ is the source strength. The room-beamformer transfer function is defined as $B(\omega) = Y(\omega)/S(\omega)$ and written $B(\omega) = \boldsymbol{w}^H(\omega)\boldsymbol{h}(\omega)$ where

$$\boldsymbol{w}(\omega) \triangleq [W_1(\omega), W_2(\omega), ..., W_N(\omega)]^H,$$

and

$$\boldsymbol{h}(\omega) \triangleq [H_1(\omega), H_2(\omega), ..., H_N(\omega)]^T.$$

We express the vector $\boldsymbol{h}(\omega)$ of room transfer functions as the sum of the direct part

$$\boldsymbol{h}_{\mathrm{d}}(\omega) = [H_1^{(\mathrm{d})}(\omega), H_2^{(\mathrm{d})}(\omega), ..., H_N^{(\mathrm{d})}(\omega)]^T,$$

and reverberant part

$$\boldsymbol{h}_{\mathrm{r}}(\omega) = [H_1^{(\mathrm{r})}(\omega), H_2^{(\mathrm{r})}(\omega), ..., H_N^{(\mathrm{r})}(\omega)]^T.$$

The elements of $\boldsymbol{h}_{\mathrm{d}}(\omega)$ are equal to the free field transfer function for an omnidirectional source [67] scaled by the radiation pattern of the source $Q(\hat{\boldsymbol{\phi}};\omega)$ in each source-to-sensor direction $\hat{\boldsymbol{\phi}}_n$:

$$H_n^{(\mathrm{d})}(\omega) = \frac{Q(\hat{\boldsymbol{\phi}}_n;\omega)}{4\pi\|\boldsymbol{y}-\boldsymbol{x}_n\|} e^{-ik\|\boldsymbol{y}-\boldsymbol{x}_n\|}, \qquad n = 1, 2, \ldots, N, \tag{6.1}$$

where $\hat{\boldsymbol{\phi}}_n \triangleq (\boldsymbol{x}_n - \boldsymbol{y})/\|\boldsymbol{x}_n - \boldsymbol{y}\|$. To ensure insertion of $Q(\hat{\boldsymbol{\phi}};\omega)$ has no impact on the source strength, $Q(\hat{\boldsymbol{\phi}};\omega)$ is normalized so that:

$$\frac{1}{4\pi}\int_{\mathbb{S}^2} |Q(\hat{\boldsymbol{\phi}};\omega)|^2 ds(\hat{\boldsymbol{\phi}}) = 1.$$

For a description of the reverberant component $\boldsymbol{h}_{\mathrm{r}}(\omega)$, we refer to the diffuse field model presented in Chapter 2. Though the sound source in the current chapter is directional, to good approximation the diffuse field model is still valid. Since reflections off walls are typically at least partially diffuse, sequential reflections of wavefronts off the walls cause the originally directional sound to rapidly diffuse throughout the room.

## 6.3 Beamformer Design

It is common in conventional beamformer design to constrain the design to capture of the direct part without distortion. We impose this constraint in the beamformer designs here:

$$\boldsymbol{w}^H(\omega)\boldsymbol{h}_{\mathrm{d}}(\omega) = 1. \tag{6.2}$$

Ideally we want to design the beamformer to maximize speech intelligibility. This design philosophy requires an objective measure of intelligibility. One measure studied by Thiele [101] is based on the observation that reflections with a delay time less than 50ms improve speech intelligibility while the rest of the reflections reduce it. By spatially selecting sound less than 50ms, or 17m, away from the sensors, we can maximize Thiele's intelligibility criterion.

Below we present the beamformer designs studied in this chapter. Section 6.3.1 presents a design that approximately maximizes Thiele's criterion. Section 6.3.2 presents a simplified beamformer design for the case that acoustic transfer functions $H_1^{(\mathrm{r})}(\omega), H_2^{(\mathrm{r})}(\omega), \ldots, H_N^{(\mathrm{r})}(\omega)$ are uncorrelated.

### 6.3.1 Minimum Farfield Power

A beamformer design method that approximately maximizes Thiele's criterion is presented in [109]. Here it is argued that the most detrimental component of the reverberation lies in the farfield. Intelligibility can be maximized by designing a beamformer that spatially discriminates against farfield reverberation. We hence minimize the output power of the beamformer to farfield reverberation subject to (6.2) [109]:

$$\min_{\boldsymbol{w}} \boldsymbol{w}^H(\omega)\boldsymbol{R}(\omega)\boldsymbol{w}(\omega) \quad \text{subject to} \quad \boldsymbol{w}(\omega)^H\boldsymbol{h}_{\mathrm{d}}(\omega) = 1,$$

where for the classical diffuse field the sensor correlation matrix $\boldsymbol{R}(\omega)$ is

$$[\boldsymbol{R}(\omega)]_{mn} = \operatorname{sinc}\left(k\|\boldsymbol{x}_m - \boldsymbol{x}_n\|\right), \qquad m, n = 1, 2, \ldots, N, \tag{6.3}$$

the term $[\boldsymbol{R}(\omega)]_{mn}$ corresponds to the element in the $m$th row and $n$th column of $\boldsymbol{R}(\omega)$ and $\|\boldsymbol{x}_m - \boldsymbol{x}_n\|$ is recognized as sensor-sensor spacing. This constrained linear optimization problem has the well known solution

$$\boldsymbol{w}_{\mathrm{ff}}(\omega) = \frac{\boldsymbol{R}^{-1}(\omega)\boldsymbol{h}_{\mathrm{d}}(\omega)}{\boldsymbol{h}_{\mathrm{d}}^H(\omega)\boldsymbol{R}^{-1}(\omega)\boldsymbol{h}_{\mathrm{d}}(\omega)}. \tag{6.4}$$

At low frequencies, implementation of the design becomes ill-conditioned and requires regularization. As $\omega \to 0$, $\boldsymbol{R}(\omega)$ turns into a matrix of 1's and hence becomes singular. This problem is avoided by diagonal loading $\boldsymbol{R}(\omega)$, that is by adding a

small number $\epsilon$ to the diagonal of $\boldsymbol{R}(\omega)$:

$$\boldsymbol{R}'(\omega) = \boldsymbol{R}(\omega) + \epsilon \boldsymbol{I}_N,$$

where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix.

### 6.3.2 Minimum White Noise Gain

For comparison, we derive a simpler design called the minimum white noise gain (WNG) beamformer [26]. This design also attempts to minimize farfield power. It does not however account for the correlation of the interfering noise between sensors, assuming it to be zero. This design minimizes the WNG, the power gain due to white noise of unity variance at the sensors:

$$\min_{\boldsymbol{w}} \boldsymbol{w}^H(\omega)\boldsymbol{w}(\omega) \quad \text{subject to} \quad \boldsymbol{w}(\omega)^H \boldsymbol{h}_{\mathrm{d}}(\omega) = 1,$$

It is equivalent to the above design in the case $\boldsymbol{R}(\omega) = \boldsymbol{I}$, such as is created at frequencies for which sensors are either well-separated ($k\|\boldsymbol{x}_m - \boldsymbol{x}_n\| \gg 1$) or spaced $\lambda/2$ apart. The solution is simply

$$\boldsymbol{w}_{\mathrm{wng}}(\omega) = \frac{1}{\boldsymbol{h}_{\mathrm{d}}^H(\omega)\boldsymbol{h}_{\mathrm{d}}(\omega)} \boldsymbol{h}_{\mathrm{d}}(\omega). \tag{6.5}$$

In terms of original parameters

$$[\boldsymbol{w}_{\mathrm{wng}}(\omega)]_n = \frac{1}{\boldsymbol{h}_{\mathrm{d}}^H(\omega)\boldsymbol{h}_{\mathrm{d}}(\omega)} \frac{Q(\hat{\boldsymbol{\phi}}_n;\omega)}{4\pi\|\boldsymbol{y} - \boldsymbol{x}_n\|} e^{ik\|\boldsymbol{y}-\boldsymbol{x}_n\|},$$

where $[\boldsymbol{w}_{\mathrm{wng}}(\omega)]_n$ is the $n$th element of $\boldsymbol{w}_{\mathrm{wng}}(\omega)$. Contrasted with these weights are the delay-and-sum beamformer weights

$$[\boldsymbol{w}_{\mathrm{ds}}(\omega)]_n = e^{ik\|\boldsymbol{y}-\boldsymbol{x}_n\|}. \tag{6.6}$$

Ignoring the frequency dependent normalization constant $1/\boldsymbol{h}_{\mathrm{d}}^H(\omega)\boldsymbol{h}_{\mathrm{d}}(\omega)$, the WNG beamformer is similar to the delay-and-sum beamformer in that it time aligns the direct parts of the sensor signals. The WNG beamformer also applies more weight to the sensors that are either in front of or closer to the source.

## 6.4 Performance Measures for Beamformers

We now describe the measures of performance of the beamformers. Section 6.4.1 defines the frequency-weighted direct-to-reverberant ratio that is used to quantify reverberation suppression. Section 6.4.2 summarizes the speech transmission index

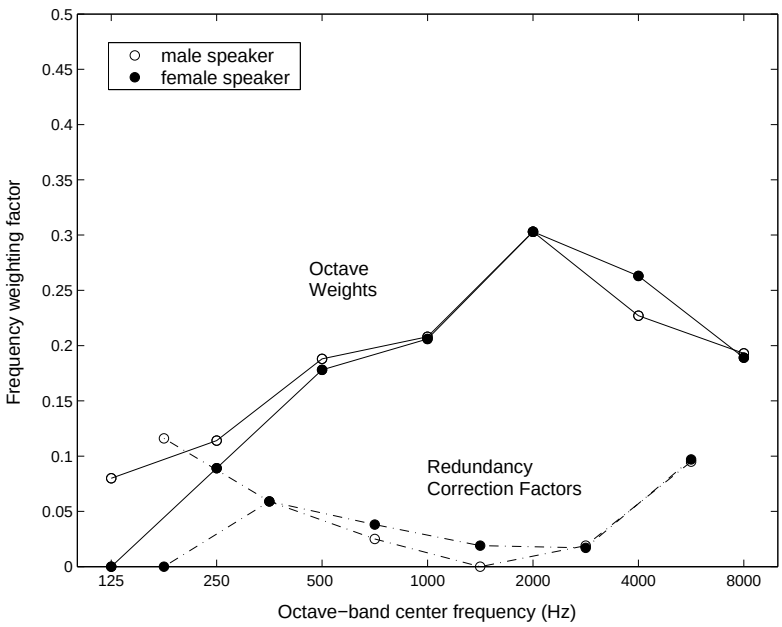



Figure 6.2: Speech transmission index octave weights (solid lines) and redundancy correction factors (broken lines) for male and female speakers. Data is taken from [90].

(STI), a far more comprehensive measure of speech intelligibility, accounting for both time-domain and frequency-domain perceptual phenomena.

These measures are applicable to quantifying the intelligibility of speech transmission in any acoustic system. They are traditionally applied to the speech captured with a microphone in a room, but below their use is extended to the speech captured with a beamformer.

### 6.4.1 Frequency Weighted Direct-to-Reverberant Ratio

We quantify the level of reverberation suppression with the frequency-weighted DRR. In this ratio, the energies of the direct and reverberant parts of the beamformer output signal are weighted over frequency:

$$\overline{\gamma}_{\mathrm{B}} \triangleq \frac{\int_{-\infty}^{\infty} |S(\omega)|^2 |B_{\mathrm{d}}(\omega)|^2 \, d\omega}{\int_{-\infty}^{\infty} |S(\omega)|^2 E\{|B_{\mathrm{r}}(\omega)|^2\} \, d\omega}. \tag{6.7}$$

The expectation operator is present because the diffuse reverberant field is modelled with random variables. Given a room-beamformer impulse response $b(t)$ and source signal spectrum $S(\omega)$, we can readily determine $B_{\mathrm{d}}(\omega)$ and $B_{\mathrm{r}}(\omega)$ and hence calculate $\overline{\gamma}_{\mathrm{B}}$.

Important to calculation of the frequency weighted DRR is the choice of $S(\omega)$. One choice is to set $S(\omega)$ equal to 1 in the frequency range 100Hz-10kHz and zero otherwise. This choice provides a measure of the reverberation suppression of the beamformer in the range of frequencies of interest to speech.

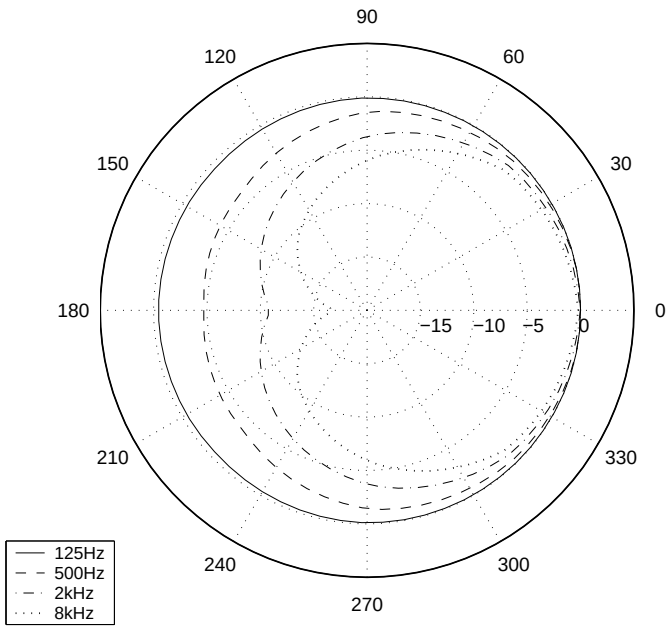



Figure 6.3: Human speaker directional response (dB) obtained by least squares fitting to [9]. The average directivity index over the 100Hz - 10kHz range is 4.4 dB.

Another choice is to set $|S(\omega)|^2$ over each octave band equal to the STI octave weights of Figure 6.2. These weights reflect the relative importance of difference frequencies to the listener[1]. The frequency weighted DRR however does not quantify the effect of syllabic blurring on intelligibility. However this perceptual phenomena is better quantified through the STI.

### 6.4.2 Speech Transmission Index

The speech transmission index (STI) is an objective measure of speech intelligibility over acoustic channels [42]. The STI is calculated through measuring the modulation transfer function (MTF). The MTF is calculated for 14 modulation frequencies at each of seven octave-bandlimited test signals, and converted into seven effective signal to noise ratios (SNRs). The STI is then calculated as a function of these 98 effective SNRs.

In the six steps below we describe the STI determination in detail:

*1) Modulation Transfer Function:* The MTF is defined as the modulation index of the intensity envelope of a transmitted test signal as a function of frequency (see Figure 6.4). 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 modulating the periodic function $p(t) = \sqrt{1 + \cos(2\pi F t)}$, where $F$ is the frequency of the modulating sinusoid. This periodic function produces a signal intensity $p^2(t)$ that is a full swing sinusoid.

[1]The weights are slightly different for male and female speakers. Lower frequencies appear to be more important to intelligibility of male speech.

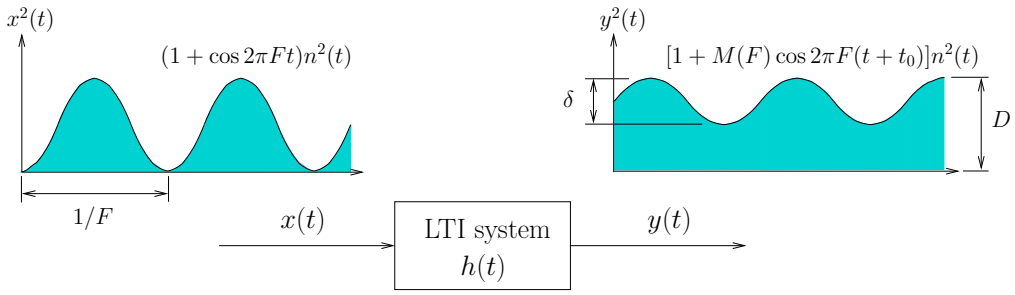



Figure 6.4: Illustration of the modulation transfer function, in the linear time invariant (LTI) system $y(t) = h(t) * x(t)$. The modulation index here is $M(f) = \delta/D$. The shading represents the intensity of the modulating white noise $n(t)$.

Over noiseless 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:

$$M(F) = \frac{\int_0^\infty h^2(t)e^{2\pi iFt}\,\mathrm{d}t}{\int_0^\infty h^2(t)\,\mathrm{d}t}, \tag{6.8}$$

where $h(t)$ is a causal impulse response. This equation is derived in [43] but we provide a clearer derivation in Appendix B.

For the MTF used in STI, the white noise is octave-bandlimited by filtering with 6th order Butterworth filters of center frequencies ranging from 125Hz to 8kHz [89]. For octave-bandlimited noise, (6.8) is still valid, provided that (i) $h(t)$ is obtained by convolving the room-beamformer impulse response with the impulse response of the bandpass filter; and (ii) the bandwidth of the bandpass filter is much larger than the modulating frequency. This fact is also shown in Appendix B.

*2) Conversion to Effective SNRs:* The STI definition requires that $M(F)$ for each octave be sampled for 14 modulation frequencies $F_m$ 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,

$$\mathrm{SNR}_n(F_m) = 10\log\left(\frac{M_n(F_m)}{1-M_n(F_m)}\right),$$

where $n = 1, 2, \ldots, 7$ and $m = 1, 2, \ldots, 14$. Subscript $n$ refers to the octave band.

*3) Range Limiting:* There exists lower and upper SNR limits outside of which the difference made to speech intelligibility is small. Modelling this effect, STI hard limits effective SNR to the range $-15$dB to 15dB. Values outside of this range are truncated to $\pm 15$ dB.

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

$$\overline{\mathrm{SNR}}_n = \frac{1}{14}\sum_{m=1}^{14}\mathrm{SNR}_n(F_m).$$

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

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

*6) Octave Weighting:* Finally, the STI is obtained by applying weighting factors $\alpha_n$ to the octave-band-specific SNRs and summing. These factors weight the perceptual effect of different frequencies to the listener. Recent findings [90] have found that contributions from different frequency bands are not purely additive. Large signal energy in one octave band tends to mask the energy in adjacent frequency bands. To account for this effect, Steeneken suggests the inclusion of the redundancy correction factor $\beta_n$:

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

The values of $\alpha_n$ and $\beta_n$ used were obtained from the results of [90], where the STI was fit to CVC word score data. These octave weights are shown in Figure 6.2 for both male and female speakers.

The difference between sexes, though minor, reminds the reader that speech perception is a complicated phenomenon. Although we do not account for it here, intelligibility is also dependent on the type of phoneme spoken [91].

## 6.5 Human Speaker Radiation Pattern

To simulate the radiation pattern of a human speaker, we least squares fitted to human speaker data obtained by Dunn and Farnsworth [24]. Presented in Figure 6.3 are the curves of best fit to the 60cm $x$-$y$ plane sound pressure data.

Although the Farnsworth data is quite noisy, it has the advantage of being based on measurement of a live human speaker. Because such experiments are so time consuming to perform, they are rarely found in the literature[2]. Most authors resort to simulation with a dummy head or mannequin with an in-built transducer [30,44] or to analytical modelling of the diffusion of the sound about the head [96]. These approaches model the basic character of the radiation pattern but lack the accuracy of real human speaker data [34].

In the next section, the speaker directivity model is used in simulating beamforming to a human speaker. Using it will allow quantifying the difference that accounting for speaker directivity has on beamformer performance.

[2]Only one similar work was found [63].

| Name | Design Criterion | Filter Weight $\boldsymbol{w}(\omega)$ |
|---|---|---|
| DS | Delay and sum | $[\boldsymbol{w}_{\mathrm{ds}}(\omega)]_n = e^{ik\|\boldsymbol{y}-\boldsymbol{x}_n\|}$ |
| DWNG | Minimize white noise gain | $\dfrac{\boldsymbol{h}_{\mathrm{d}}(\omega)}{\boldsymbol{h}_{\mathrm{d}}^H(\omega)\boldsymbol{h}_{\mathrm{d}}(\omega)}$ |
| NWNG | Miminize white noise gain but assume $Q(\hat{\boldsymbol{\phi}};\omega) \equiv 1$ | $\dfrac{\boldsymbol{h}_{\mathrm{d}}(\omega)}{\boldsymbol{h}_{\mathrm{d}}^H(\omega)\boldsymbol{h}_{\mathrm{d}}(\omega)}$ |

Table 6.1: Beamformer designs used in simulations.

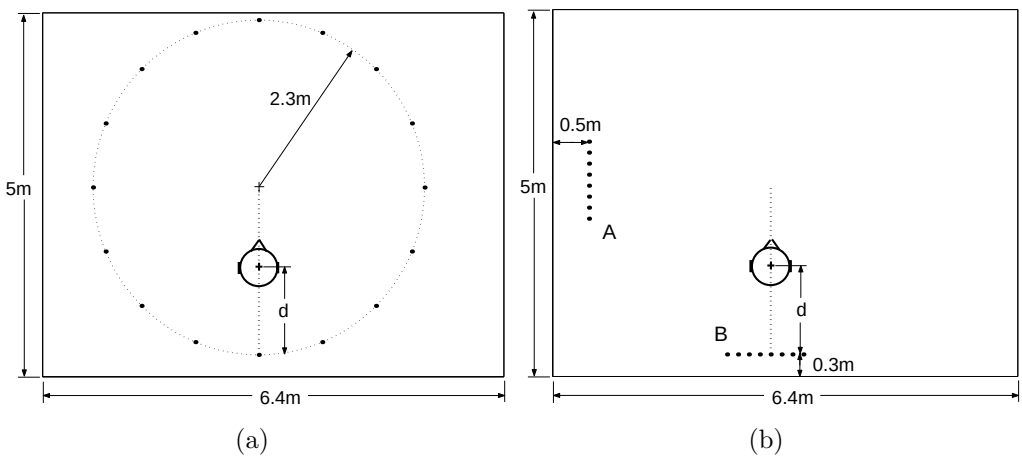



Figure 6.5: The room setup for (a) the circular array simulation, (b) the simulation with pair of linear arrays. The speaker faces north and moves along the vertical dotted line. Each sensor is shown as a dot (·).

## 6.6 Simulation

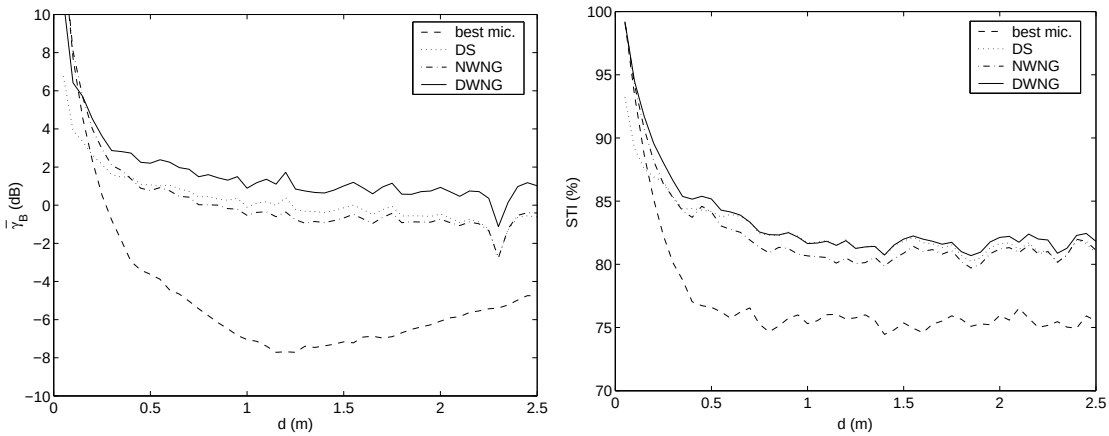
We now simulate sound capture with the above beamformer designs in a reverberant room, to quantify the improvements obtainable with directional-source beamforming.

The beamformer designs we compare in simulation are summarized in Table 6.1. The minimum farfield power beamformer is excluded from the simulation results, since its performance we observed in simulation to be very close to the WNG beamformer. For comparison we also include the best sensor reading.

The reverberant room setups used in simulation are summarized in Figure 6.5. We simulated reverberation in a rectangular room of dimensions $6.4 \times 5 \times 4$m and $c = 342$m/s, using the image-source model [4]. The source and the sensors were placed 2.5m above the floor of the room.

The radiation pattern of the source has been accounted for in the image-source model by assuming a directional source and omnidirectional images. Though approximate, this simplified treatment is appropriate in our simulations since the source at most frequencies has only a modest directional gain and the room has equal wall absorption coefficients[3].

[3]Only in extreme cases does this model break down. One example is a highly directional




(a) Frequency weighted DRR (b) STI

Figure 6.6: Results of the beamformer simulations for the uniform circular array.

We perform simulations on (i) a uniform circular array (Section 6.6.1), (ii) pair of uniform linear arrays (Section 6.6.2) and (iii) single uniform linear array in a highly reverberant field (Section 6.6.3). Tests (i) and (ii) allow comparison of the directional WNG beamformer with the other techniques. Since test (iii) allows conclusions to be drawn on reverberant room beamforming in general. For further comparison, we calculate the frequency weighted DRR for the classical diffuse reverberant field (Section 6.6.4).

In each case, we measure the frequency weighted DRR and STI. Frequency weighted DRRs were calculated with equal weighting over the 100Hz - 10kHz frequency range. STI has been computed using the frequency weights for a female talker.

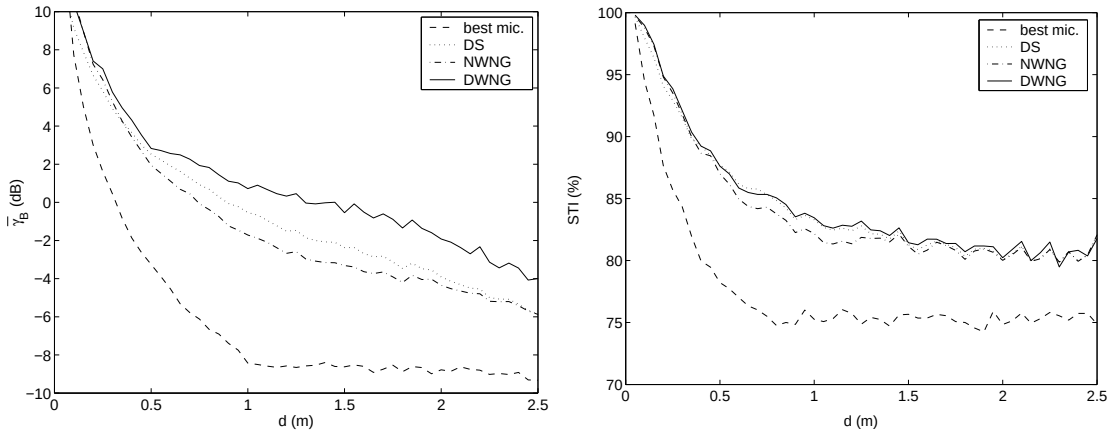
### 6.6.1 Circular Array

Simulation is performed on the 16-element uniform circular array of radius 2.2m centered at $(3, 2.6, 2.5)$ (Figure 6.5(a)), in a room with wall absorption coefficient of 0.2 and reverberation time of 700ms. This example represents a best case scenario for directional source beamforming. Provided the source lies in the interior of the array, no matter which direction the source faces, sensors will always lie in front of it.

In Figure 6.6(a), directional WNG beamformer outperforms the other beamformers at all but very small source-sensor distances[4]. This beamformer outperforms other designs by up to 1.8dB. The directional WNG beamformer also yields best speech intelligibility (Figure 6.6(b)), beating other beamformer methods by

source pointing into the only perfectly absorbing wall of a room: all sound is absorbed by the first reflection and this model will fail to yield an accurate description of the reverberation.

[4]Because of the unity-gain constraint, the directional WNG beamformer compensates for the low frequency roll-off caused by positioning a sensor behind the source, by amplifying high frequencies. This compensation degrades the DRR.




(a) Frequency weighted DRR (b) STI

Figure 6.7: Results of the beamformer simulations for the pair of uniform linear arrays.

upto 2% STI. Unfortunately a 2% STI improvement is barely audible. A more noticeable intelligibility improvement is the $7 - 10\%$ STI of using any particular beamformer design over a single microphone.

### 6.6.2 Pair of Linear Arrays

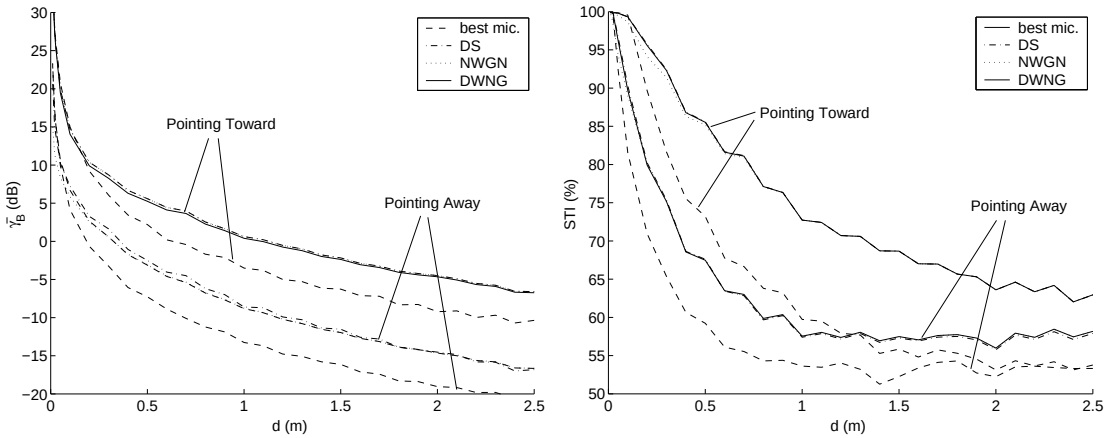
Simulation is performed on the pair of 8-element uniform linear arrays with 0.15m a sensor-sensor spacing (see Figure 6.5), in a room with wall absorption coefficient of 0.2. For this geometry, directional source beamforming best enhances beamformer performance when the source is facing away from one of the linear arrays. We consider this case here.

In Figure 6.7(a) we see that the directional WNG beamformer significantly outperforms all other schemes in frequency weighted DRR by up to 2dB. The other beamformers tend to perform worse, as these designs apply too much weight to the sensors positioned behind the source. However in Figure 6.7(b) we see that in STI, all beamformers perform about the same.

### 6.6.3 Linear Array in Strong Reverberation

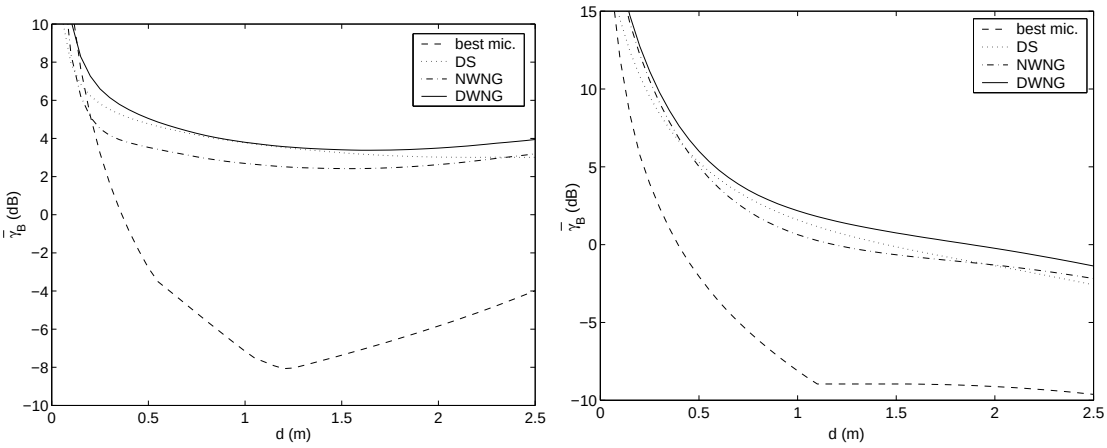
We now present a simulation that allows a study of the intelligibility improvement that is relatively independent of the beamformer design. Simulation is performed on a 8-element uniform linear array with a 0.15m sensor-sensor spacing (array B in Figure 6.5), for the speaker first facing the beamformer (south) and then facing away from the beamformer (north). The room has a wall absorption coefficient of 0.1 and reverberation time of 1.3sec.

In Figure 6.8(a), the beamforming performs up to 5dB better in DRR than the best sensor. However it performs up to 10dB better by pointing the source at the sensor array, and much better by simply locating the source next to the sensor




(a) Frequency weighted DRR (b) STI

Figure 6.8: Results of the beamformer simulations for a uniform linear array in a highly reverberant field. The dotted line represents the the best sensor reading. Results are plotted for the speaker facing both toward and away from the array.




(a) Uniform circular array (b) Pair of uniform linear arrays

Figure 6.9: Diffuse field DRR for the uniform circular array and pair of uniform linear arrays.

array. In Figure 6.8(b) the results are similar, but the STI improvements obtained by locating the sensor array in front of the source are .

### 6.6.4 Diffuse Field

For further comparison, the frequency weighted DRR is calculated for the case that the reverberation is modelled by a diffuse field. Here, the term in (6.7) involving the room-beamformer transfer function $B_{\mathrm{r}}(\omega) = \boldsymbol{w}^H(\omega)\boldsymbol{h}_{\mathrm{r}}(\omega)$ can be written:

$$E\{|B_{\mathrm{r}}(\omega)|^2\} = E\{|H_n^{(\mathrm{r})}(\omega)|^2\}\boldsymbol{w}^H(\omega)\boldsymbol{R}(\omega)\boldsymbol{w}(\omega), \tag{6.9}$$

where $\boldsymbol{R}(\omega) = E\{\boldsymbol{h}_{\mathrm{r}}(\omega)\boldsymbol{h}_{\mathrm{r}}^H(\omega)\}/E\{|H_n^{(\mathrm{r})}(\omega)|^2\}$ is the sensor correlation matrix. $E\{|H_n^{(\mathrm{r})}(\omega)|^2\}$ is proportional to mean square pressure, and is hence constant across all sensors. For a diffuse field, $\boldsymbol{R}(\omega)$ is given by (6.3) and $E\{|H_n^{(\mathrm{r})}(\omega)|^2\}$ is given

by [7, p. 312 - 314]:

$$E\{|H_n^{(\mathrm{r})}(\omega)|^2\} = \frac{4\rho_0 c}{\mathrm{SA}} \frac{1-\overline{\alpha}}{\overline{\alpha}} \Pi(\omega),$$

$\rho_0$ is the density in air, SA is wall surface area, $\overline{\alpha}$ is the average wall absorption coefficient and $\Pi(\omega)$ is source power output, related to source strength by $\Pi(\omega) = \rho_0 c k^2 |S(\omega)|^2/8\pi$ [114, p. 199].

The frequency weighted DRR for the diffuse field case has been plotted for both array geometries (Figures 6.9). These plots preserve the general trends of Figures 6.6(a) and 6.7(a).

One result of source directionality is frequency distortion in the captured sound signal. For the delay-and-sum beamformer, the direct part $\boldsymbol{w}^H(\omega)\boldsymbol{h}_{\mathrm{d}}(\omega)$ is distorted by factor $\sum_{n=1}^{N} \frac{Q(\hat{\boldsymbol{\phi}}_n;\omega)}{\|\boldsymbol{y}-\boldsymbol{x}_n\|}$. For a human speaker, whose radiation pattern shown in Figure 6.3, this distortion is a frequency roll-off. The largest roll-off ($-2.5$dB / octave) is seen for an array positioned directly behind the speaker. This lowpass filtering of the speech signal is compensated for in the directional WNG beamformer designs.

### 6.6.5 Conclusions

The difference made in beamformer designs by accounting for source directionality is significant to reverberation suppression but not intelligibility. In contrast, the intelligibility improvements for simply locating the sensors in front of or close to the source are much more significant than that due to the beamformer gain.

The idea that specifically designing to minimize farfield beamformer power improves intelligibility is not supported by the trends observed in this study. In every case considered, the non-directional WNG beamformer is outperformed in STI by the delay-and-sum beamformer. This observation is supported by [80], where the ability of beamforming for range discrimination is typically shown to be small[5].

## 6.7 Summary and Contribution

This chapter examined the improvement that can be attained by exploiting perfect knowledge of the sound source directionality in beamformer designs. This work has highlighted the utility of objective measure of speech intelligibility to assess beamformer performance.

The chapter contribution can be itemized as follows:

i. Exploration of the improvement to beamformer designs by exploiting perfect

[5] In [80], for a uniform linear array with $\lambda/2$ sensor spacing, unless the source is closer than one quarter of the array length to the sensors, beamformer gain is improved by no more than 1dB.

knowledge of the source radiation pattern and orientation. This improvement was quantified in several key cases.

ii. Study of the ability of the minimum farfield power and minimum white noise gain beamformers to suppress the reverberation that is detrimental to speech intelligibility.

iii. Application of an objective speech intelligibility measure, the STI, to assess performance of beamformer designs.