

Loudspeaker Beamforming to Enhance Speech Recognition Performance of Voice Driven Applications

Groot, D. de; Karslioglu, B.; Scharenborg, O.; Martinez, J.

DOI

[10.1109/icassp49660.2025.10889702](https://doi.org/10.1109/icassp49660.2025.10889702)

Publication date

2025

Document Version

Final published version

Published in

ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Citation (APA)

Groot, D. D., Karslioglu, B., Scharenborg, O., & Martinez, J. (2025). Loudspeaker Beamforming to Enhance Speech Recognition Performance of Voice Driven Applications. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings).
<https://doi.org/10.1109/icassp49660.2025.10889702>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

**Green Open Access added to [TU Delft Institutional Repository](#)
as part of the Taverne amendment.**

More information about this copyright law amendment
can be found at <https://www.openaccess.nl>.

Otherwise as indicated in the copyright section:
the publisher is the copyright holder of this work and the
author uses the Dutch legislation to make this work public.

Loudspeaker Beamforming to Enhance Speech Recognition Performance of Voice Driven Applications

D. de Groot, B. Karslioglu, O. Scharenborg, J. Martinez

Multimedia Computing Group, EEMCS, Delft University of Technology, The Netherlands

{d.c.c.j.degroot, j.a.martinezcastaneda, o.e.scharenborg}@tudelft.nl, b.karslioglu@student.tudelft.nl

Abstract—In this paper we propose a robust loudspeaker beamforming algorithm which is used to enhance the performance of voice driven applications in scenarios where the loudspeakers introduce the majority of the noise, e.g. when music is playing loudly. The loudspeaker beamformer modifies the loudspeaker playback signals to create a low-acoustic-energy region around the device that implements automatic speech recognition for a voice driven application (VDA). The algorithm utilises a distortion measure based on human auditory perception to limit the distortion perceived by human listeners. Simulations and real-world experiments show that the proposed loudspeaker beamformer improves the speech recognition performance in all tested scenarios. Moreover, the algorithm allows to further reduce the acoustic energy around the VDA device at the expense of reduced objective audio quality at the listener's location.

Index Terms—Spotforming, beamforming, speech recognition

I. INTRODUCTION

AUTOMATIC speech recognition (ASR) is steadily being integrated into our daily lives. While ASRs can attain high accuracy in clean acoustic conditions, they are often deployed as part of voice driven applications (VDAs), such as voice assistants like *Amazon Alexa*, *Apple Siri* and *Google Assistant*. The scenarios where VDAs should work include noisy and reverberant environments, e.g. living rooms and cars [1], [2], [3]. These environments pose a significant challenge to perform ASR [4], as the algorithms are usually not trained on different types of signal degradations but only on clean speech [5], [6]. In practice different types of speech enhancement techniques, e.g. denoising and microphone beamforming, are used as preprocessing steps to improve ASR performance [7], [8], [9]. Yet, ASR performance in adverse acoustic conditions is not always satisfactory [10]. This has prompted researchers to use a large amount of microphones [11], or to consider different sensor modalities such as video [6], bone-conduction microphones [12] and radar [13]. While effective, these systems rely on specialized (sometimes intrusive) hardware or to record video, which is not always feasible or desirable. In this work, we propose a loudspeaker beamforming algorithm to improve ASR performance in VDA systems with multiple loudspeakers and where the loudspeaker signals are themselves a major source of interference. This is the case, for example, when the user is watching a movie in their home cinema and wants to pause the movie with a voice command or when the user likewise wants to pause music in their car. We are not the first to consider the loudspeaker modality to enhance ASR performance. In [14], a method combining sound field synthesis and acoustic echo cancellation (AEC) was proposed. Compared to our method, this method has more stringent requirements on the number of loudspeakers. Additionally, in order to perform AEC, it was assumed that the playback signals are available at the VDA. This assumption is not satisfied when only a low capacity link is available between the playback system and the VDA. Although AEC is not included in this work, our algorithm allows for its integration if the playback signals are available to the VDA.

The proposed loudspeaker beamforming algorithm is based on a class of robust near-field microphone beamformers called microphone spotformers, in which the signal from some region-of-interest is amplified or attenuated [15], [16]. We adapt the microphone spotformer to create a “loudspeaker spotformer” (LSp). The proposed system

modifies the loudspeaker playback signals such that a low-acoustic-energy region is formed around the VDA while maintaining a high reproduction quality with low perceived distortion around the user. Keeping the perceived distortion limited is challenging in practice. For this, an objective measure of human sound perception well-suited for optimisation is required. This is an active research topic in e.g. sound zone synthesis [17], [18]. We constrain the distortion introduced to the listener using a measure of human auditory masking [19], as in [18]. The proposed LSp algorithm assumes that the location of the loudspeakers, the user, and the VDA lie within regions that can be estimated or measured a priori. This is a reasonable assumption: in a car or a living room the locations of the loudspeakers, the user, and the VDA mostly remain within a priori established regions. Our proposed algorithm is evaluated through simulations and in a real-world scenario. The Matlab code implementing the loudspeaker spotformer can be found at [20].

II. NOTATION AND SIGNAL MODEL

Consider a space in which L loudspeakers are placed at positions $\mathbf{x}_L^{(l)} \in \mathbb{R}^3$, $l \in \{1, \dots, L\}$. For each loudspeaker, we define $s_{\text{ref}}^{(l)}$ as the reference loudspeaker playback signal (no LSp algorithm applied). Likewise, $s_L^{(l)}$ is the loudspeaker playback signal when the LSp algorithm is applied. In the room there is a VDA equipped with a microphone array implementing microphone beamforming. The center of the array is located at $\mathbf{x}_M \in \mathbb{R}^3$. In practice, the locations of the VDA and loudspeakers can be measured by hand or estimated through signal processing techniques such as [21]. There is a listener assumed to be approximately at a known location $\mathbf{x}_u \in \mathbb{R}^3$. We use the free-field acoustic transfer function from the loudspeakers to P points in the neighbourhood of \mathbf{x}_u as control points: $\mathbf{x}_p^{(p)} \in \mathbb{R}^3$, $p \in \{1, \dots, P\}$. The room impulse response (RIR) from location \mathbf{x}_s to location \mathbf{x}_r is given by $h(\mathbf{x}_s, \mathbf{x}_r, t)$. Frequency-domain variables are indicated with a hat on top of the corresponding symbol. The room transfer function (RTF) is given by $\hat{h}(\mathbf{x}_s, \mathbf{x}_r, \omega)$, with $\omega = 2\pi f$ and f the frequency in hertz. We model each RTF as consisting of the direct path only, eliminating the need to estimate the highly varying RTF. Instead we rely on the robustness of the LSp algorithm. Thus, the RTF is given by

$$\hat{h}(\mathbf{x}_s, \mathbf{x}_r, \omega) = \frac{e^{-j\omega \|\mathbf{x}_s - \mathbf{x}_r\|_2 / c}}{4\pi \|\mathbf{x}_s - \mathbf{x}_r\|_2}, \quad (1)$$

with $j^2 = -1$ and $c = 342$ m/s the sound velocity [16], [22].

III. LOUDSPEAKER SPOTFORMING FOR VDAs

In this section the setup used for our simulations and real-world test is introduced and the loudspeaker spotformer (LSp) is derived. In Fig. 1 the setup is depicted. The left hand side (Fig. 1a) provides a top-view schematic of the room with the position of the loudspeakers $\mathbf{x}_L^{(l)}$, the position of the VDA featuring a circular microphone array (region \mathcal{M}), and the position of the listener zone depicting the control points $\mathbf{x}_p^{(p)}$. To the right (Fig. 1b), a picture of the implemented setup in a real room is given. A zoom-in picture featuring our implementation of the VDA with microphone array is shown on the top right. In Fig. 1a the microphones are depicted by dots (\bullet) and

the control points by crosses (\times). Notice that region \mathcal{M} (in red) is not a circle but a torus. Mathematically the torus reflects the spatial distribution of the circular microphone array more accurately.

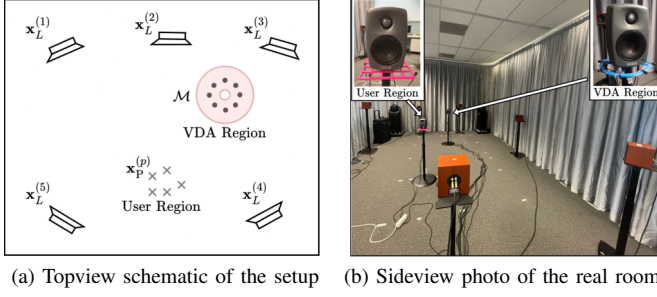


Fig. 1. The loudspeaker spotformer (LSp) setup. In (a), a topview schematic of the setup is shown. The LSp computes the loudspeaker playback signals which minimise the acoustic energy in region \mathcal{M} around the microphones (\bullet) of the voice driven application (VDA). The control points (\times) are not physically placed but modelled within the region the user is expected to be listening. The algorithm limits the acoustic distortion at these points. In (b), a photo of the actual experiment setup is shown. A zoomed-in photo of our VDA implementation using a circular microphone array is shown in the top right corner. In the top-left corner, a zoom-in photo of the loudspeaker emulating the user in the experiments is shown. The microphones on the pink grid are used to evaluate the audio quality in Sec. IV-B.

In the following we compute the spatial covariance matrix $\mathbf{R}_{\mathcal{M}}$ describing the acoustic energy in region \mathcal{M} . The LSp algorithm is obtained by combining this covariance matrix with a distortion measure based on human auditory masking.

A. Spatial Covariance Matrix

Define the loudspeaker-to-receiver transfer vector $\hat{\mathbf{v}}_L \in \mathbb{C}^L$ as

$$\hat{\mathbf{v}}_L(\mathbf{x}_r, \omega) = [\hat{h}(\mathbf{x}_r, \mathbf{x}_L^{(1)}, \omega), \dots, \hat{h}(\mathbf{x}_r, \mathbf{x}_L^{(L)}, \omega)]^T, \quad (2)$$

and the vector of frequency-domain loudspeaker signals $\hat{\mathbf{s}}_L$ as

$$\hat{\mathbf{s}}_L(\omega) = [\hat{s}_L^{(1)}(\omega), \dots, \hat{s}_L^{(L)}(\omega)]^T. \quad (3)$$

The audio received at \mathbf{x}_r is given by $\hat{y}_r(\mathbf{x}_r, \omega) = \hat{\mathbf{s}}_L^T(\omega) \hat{\mathbf{v}}_L(\mathbf{x}_r, \omega)$. Now assume that \mathbf{x}_r is a realisation of random vector $\mathbf{x}_{\mathcal{M}}$ with corresponding distribution $p_{\mathcal{M}}(\mathbf{x}_{\mathcal{M}})$, which can be interpreted as representing the probability of finding the microphones of the VDA over some spatial region. This spatial stochastic model is the key insight that provides the algorithm its robustness against errors in the positions. We return to this distribution later. Given random vector $\mathbf{x}_{\mathcal{M}}$, the signal received in region \mathcal{M} is given by

$$y_{\mathcal{M}}(\mathbf{x}_{\mathcal{M}}, \omega) = \hat{\mathbf{s}}_L^T(\omega) \hat{\mathbf{v}}_L(\mathbf{x}_{\mathcal{M}}, \omega). \quad (4)$$

The expected value of the acoustic energy in region \mathcal{M} is

$$\begin{aligned} \mathbb{E} \{ |y_{\mathcal{M}}(\mathbf{x}_{\mathcal{M}}, \omega)|^2 \} &= \hat{\mathbf{s}}_L^T(\omega) \mathbf{R}_{\mathcal{M}}(\omega) \hat{\mathbf{s}}_L^*(\omega) \\ &= \left\| \mathbf{L}_{\mathcal{M}}^H(\omega) \hat{\mathbf{s}}_L^*(\omega) \right\|_2^2, \end{aligned} \quad (5)$$

with $\mathbb{E} \{ \hat{\mathbf{v}}_L(\mathbf{x}_{\mathcal{M}}, \omega) \hat{\mathbf{v}}_L^H(\mathbf{x}_{\mathcal{M}}, \omega) \} = \mathbf{R}_{\mathcal{M}}(\omega) = \mathbf{L}_{\mathcal{M}}(\omega) \mathbf{L}_{\mathcal{M}}^H(\omega)$. Matrix $\mathbf{L}_{\mathcal{M}}(\omega)$ follows from the Cholesky factorisation of positive-semidefinite matrix $\mathbf{R}_{\mathcal{M}}(\omega)$ [23], [24, Cor. 7.2.9]. The elements of $\mathbf{R}_{\mathcal{M}}(\omega)$ are found by spatially integrating over $p_{\mathcal{M}}$,

$$\{\mathbf{R}_{\mathcal{M}}(\omega)\}_{ll'} = \int_{\mathbb{R}^3} \hat{h}(\mathbf{x}, \mathbf{x}_L^{(l)}, \omega) \hat{h}^*(\mathbf{x}, \mathbf{x}_L^{(l')}, \omega) p_{\mathcal{M}}(\mathbf{x}) d\mathbf{x}. \quad (6)$$

Matrix $\mathbf{R}_{\mathcal{M}}$ can be augmented to account for the contribution of late reverberation by making $\mathbf{R}_M = \mathbf{R}_{\mathcal{M}} + \mathbf{R}_{\text{iso}}$, where \mathbf{R}_{iso} models the contribution of late reverberation as an isotropic acoustic field, the expression of which is known analytically [16], [25]. We return to distribution $p_{\mathcal{M}}$. This distribution can be interpreted as describing the probability of finding the circular microphone array of the VDA within

some spatial region. We express the distribution in cylindrical coordinates with radius $r \in [0, \infty)$, azimuthal angle $\theta \in [0, 2\pi)$ and height $z \in \mathbb{R}$. Due to the circular microphone array, the distribution is chosen to resemble a torus centered at \mathbf{x}_M . This is achieved by using a uniform distribution in θ and Gaussian distributions in z and r . The latter is truncated such that $r \geq 0$ [26, p. 156]. The total distribution then is

$$p_{\mathcal{M}}(r, \theta, z) = \frac{C_r}{4\pi^2 \sigma_r \sigma_z} e^{-\frac{1}{2} \left(\frac{(z - \mu_z)^2}{\sigma_z^2} + \frac{(r - \mu_r)^2}{\sigma_r^2} \right)}, \quad (7)$$

with $r \geq 0$, $\theta \in [0, 2\pi)$, C_r a normalising constant following from the truncated distribution and μ_r and μ_z the radius and height of the microphone array. The standard deviations σ_r and σ_z can be chosen based on the assumed measurement accuracy.

B. The loudspeaker spotformer

Using the covariance matrix we can minimise the acoustic energy in region \mathcal{M} . However, to avoid audible artefacts and to prevent the algorithm to converge to a trivial solution (i.e. turning all the loudspeakers off), it is important to constrain the algorithm in a smart way. We use a computationally inexpensive perceptual distortion measure based on tonal masking [19] which has been used in sound zone synthesis [18]. The measure operates on short-time frames and predicts if a distortion $\hat{\epsilon}$ is noticeable in the presence of an audible sound (called the masker) $\hat{\mathbf{s}}$. The distortion measure is given by $D(\hat{\mathbf{s}}, \hat{\epsilon}) = \|\mathbf{P}_s \hat{\epsilon}\|_2^2$ [19], with \mathbf{P}_s is a diagonal matrix based on $\hat{\mathbf{s}}$ defining a frequency dependent weighting of the distortion. See [19] for details on the computation of \mathbf{P}_s . We define each masker as the audio signal at each control point when all loudspeakers play their reference signal $\mathbf{s}_{\text{ref}}^{(l)}$. The disturbance is the deviation from the masker.

From here on we switch the notation to discrete time-domain and discrete frequency-domain for the actual implementation of the algorithm. We use a frame-length N_t and zero-padding length N_{pad} so that the frequency-domain frame length is $N = N_t + N_{\text{pad}}$. Define the time-domain playback signal vector $\mathbf{s}_L^{(l)} = [s^{(l)}(t_1) \dots s^{(l)}(t_{N_t})]^T$. The matrix of playback signals \mathbf{S}_L is

$$\mathbf{S}_L = [\mathbf{s}_L^{(1)} \dots \mathbf{s}_L^{(L)}] \in \mathbb{R}^{N_t \times L}. \quad (8)$$

The frequency-domain matrix of playback signals $\hat{\mathbf{S}}_L$ is

$$\hat{\mathbf{S}}_L = \mathbf{F} [\mathbf{I} \quad \mathbf{0}]^T \mathbf{S}_L = \mathbf{W} \mathbf{S}_L \in \mathbb{C}^{N \times L} \quad (9)$$

where \mathbf{F} is the $N \times N$ discrete Fourier transform matrix, \mathbf{I} the $N_t \times N_t$ identity matrix and $\mathbf{0}$ an $N_t \times N_{\text{pad}}$ all-zero matrix. Therefore, $\mathbf{W} \in \mathbb{C}^{N \times N_t}$ defines a zero-padded discrete Fourier transform. The time- and frequency-domain matrices of reference signals \mathbf{S}_{ref} and $\hat{\mathbf{S}}_{\text{ref}}$ are defined in the same manner. Define the loudspeaker-to-control-point transfer matrix $\mathbf{V}_p^{(l,p)} \in \mathbb{C}^{N \times N}$ as

$$\mathbf{V}_p^{(l,p)} = \text{diag} \left[\hat{h}(\mathbf{x}_L^{(l)}, \mathbf{x}_p^{(p)}, \omega_0), \dots, \hat{h}(\mathbf{x}_L^{(l)}, \mathbf{x}_p^{(p)}, \omega_{N-1}) \right]^T. \quad (10)$$

Let $\{\mathbf{A}\}_k$ be an operator that extracts the k^{th} column of \mathbf{A} . The masker $\hat{\mathbf{s}}_{\text{ref}}^{(p)}$ for control point p is given by

$$\hat{\mathbf{s}}_{\text{ref}}^{(p)} = \sum_{l=1}^L \mathbf{V}_p^{(l,p)} \{\hat{\mathbf{S}}_{\text{ref}}^T\}_l, \quad (11)$$

with corresponding masking-matrix $\mathbf{P}_{\text{ref}}^{(p)}$ as defined in [19]. The LSp is given as a convex optimisation problem [27] where the distortion at the control points is constrained. Using (5), (9), (10), and (11) this gives

$$\begin{aligned} \min_{\hat{\mathbf{S}}_L} \sum_{k=0}^{N-1} \alpha(\omega_k) \left\| \mathbf{L}_M^H(\omega_k) \{\hat{\mathbf{S}}_L^T\}_k \right\|_2^2 \\ \text{s.t. } \hat{\mathbf{S}}_L = \mathbf{W} \mathbf{S}_L, \quad \mathbf{S}_L \in \mathbb{R}^{N_t \times L}, \\ \left\| \mathbf{P}_{\text{ref}}^{(p)} \left(\sum_{l=1}^L \mathbf{V}_p^{(l,p)} \{\hat{\mathbf{S}}_L\}_l - \hat{\mathbf{s}}_{\text{ref}}^{(p)} \right) \right\|_2^2 \leq d, \quad \forall p, \end{aligned} \quad (12)$$

where the k in ω_k represents the k^{th} frequency bin, d is a parameter controlling the maximum allowable distortion calibrated to $d = 1$ when the distortion is just noticeable and $\alpha(\omega_k)$ a user-defined weighting term (see Section IV-A).

IV. EXPERIMENTAL RESULTS

In this section, the performance of the proposed LSp is evaluated. To the best of our knowledge, this is the first time loudspeaker beamforming is used to enhance voice driven applications (VDAs). Our main experiments are given in Sec. IV-C, where we test how the addition of LSp affects the ASR performance of a VDA. The word error rate (WER) and word information lost (WIL) are used as metrics for testing the ASR performance. Before this, in Sec. IV-B, we analyse the ability of the LSp to create a low-acoustic-energy region around the VDA while keeping a high reproduction quality with low perceived distortion at the user position. This is important, as our main hypothesis is that if this is correctly achieved an increase in ASR speech recognition performance follows straightforwardly. The experimental setup is described in Sec. IV-A. All experiments are carried out through simulations and by implementation in a real-world environment. The latter is important to test the claimed robustness and actual applicability of the algorithm.

A. Experimental setup and algorithm parameterisation

As mentioned in Sec. I our target application consists of systems with loudspeakers and a VDA like a voice assistant, where the loudspeaker signals are themselves a major source of interference for ASR. We choose our setup as to roughly represent a home cinema in a living room where the users are watching a movie and one would want to give a voice command to e.g. pause the movie. In the experiments we use five loudspeakers and consider a user giving voice commands. Current VDA devices are often equipped with a microphone array to enhance ASR performance. Therefore, our VDA implementation has a circular microphone array with eight microphones. The experiment setup used in both simulations and real-world experiments is shown in Fig. 2. For a photo of the real room and the VDA, see Fig. 1b.

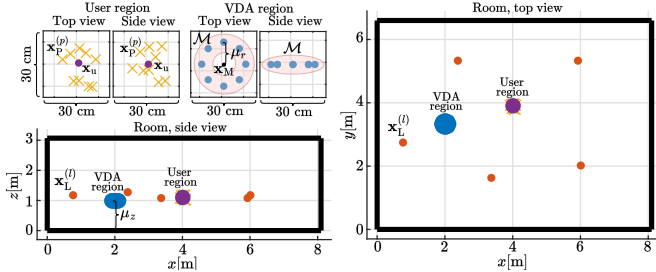


Fig. 2. A schematic view of the setup used in both simulations and real-world experiments. Zoom-ins on the user region (including control points $\mathbf{x}_p^{(p)}$ and the user location \mathbf{x}_u) and on the VDA are provided. The microphone array of the VDA has a radius $\mu_r = 0.1$ m and a height $\mu_z = 0.99$ m. Region \mathcal{M} is placed approximately on top of these microphones and is centered at the VDA location \mathbf{x}_M . The loudspeakers are located at positions $\mathbf{x}_L^{(l)}$.

In our target applications, the system would be in cases used to play music, or to make a phone-call (or video-call). For our tests we choose the loudspeaker playback signals to be segments of: an instrumental rock song, a female-voiced jazz song, a male-voiced pop song, a female-voiced speech signal and a male-voiced speech signal. Thus, the test signals consist of 40% speech, 40% vocal music and 20% instrumental music. To evaluate the ability of the LSp to create a low-acoustic-energy region around the VDA we use white Gaussian noise due to its broadband characteristic. Since the spectral content of speech is limited above 7 kHz [29] and to reduce computational complexity, a sample-rate of 16 kHz is used.

Our LSp algorithm is given by optimisation problem (12) and is solved for the loudspeaker playback signals using CVX [28]. Eq. (12) is formed through equations (5), (6), (7), (9), (10), and (11). We now describe the parameters used for these equations. For (7), which mathematically describes region \mathcal{M} (see Fig. 2) we use $\mu_r = 0.1$

m (the radius of the microphone array), $\mu_z = 0.99$ m (the height of the microphone array) and $3\sigma_r = 3\sigma_z = 0.095$ m. The number of modelled control points is $P = 9$ placed in space sampled from a normal distribution centered at the user (\mathbf{x}_u) with a standard deviation of $3\sigma = 0.2$ m. Since most important information for recognising speech starts at a frequency of approximately 100 Hz ($\omega = 200\pi$ rad/s) [29, Ch. 7.2.1], the regularisation parameter $\alpha(\omega_k) = 0$ in (12) equals 0 if $|\omega_k| \leq 200\pi$ rad/s and $\alpha(\omega_k) = 1$ otherwise. As in [30], we set the corresponding elements of matrix $\mathbf{P}_{\text{ref}}^{(p)}$ to high values (to 100 times the maximum of the inverse masking curve of the segment), encouraging the algorithm to keep the frequency content in this range unchanged. Audio is processed in segments of 16 ms (zero-padded to 32 ms) using 16 ms square-root Hanning window for analysis and synthesis with 8 ms hop length [31]. We solve (6) numerically. We select a Clenshaw-Curtis quadrature method for a good compromise between numerical complexity and accuracy [32]. For the real-world experiments, an RME Fireface UFX+ audio interface is used. We emulated the VDA in the real-world experiments using a microphone array and a loudspeaker (top-right of Fig. 1b). The microphones are AKG C417 PP lavalier microphones and the loudspeaker is a Genelec 8010AP studio monitor. The loudspeakers of the playback system are Auratone 5C Super Sound Cubes. The ‘user’ giving the voice commands is emulated by another Genelec 8010AP studio monitor (top-left of Fig. 1b). The eight microphones in the pink grid are used as validation points in the evaluation (see Section IV-B). To incorporate measurement inaccuracies in the simulated results we used 100 runs in which the loudspeaker locations deviate from their expected locations with a standard deviation of 5 cm and where the microphones of the VDA are rotated on the azimuthal plane with a standard deviation of 5° .

B. Stand-alone loudspeaker spotformer performance test

In this section we analyse the ability of the LSp to create a low-acoustic-energy region around the VDA while keeping a low perceived distortion at the user position. This is evaluated in the following scenarios:

- 1) An anechoic simulated scenario.
- 2) A simulated scenario with reverberation added using the mirror image source method (MISM) [33] as implemented by Habets [34]. The reverberation time $T_{60} \approx 220$ ms.
- 3) A real room with a reverberation time $T_{60} \approx 220$ ms.

To evaluate the distortion introduced by the LSp to the listener, the objective audio quality around the listener is measured as a function of the distortion parameter d using ViSQOLAudio [35], [36]. The loudspeaker signals used are our standard set of test signals, as named at the end of Section IV-A. The 100 simulated validation points (not to be confused with the control points $\mathbf{x}_p^{(p)}$) were sampled from a normal distribution centered at the user \mathbf{x}_u with a standard deviation of 20 cm. The 32 validation points in the real-world experiments were taken using the 18×18 cm pink grid (top-left Fig. 1b) around the listener location at four different heights in steps of 5 cm. The results are plotted in Fig. 3a as a mean opinion score (MOS) as a function of the allowed distortion d , where a MOS of 1 means bad perceived audio quality and 5 means excellent [37].

To evaluate the LSp energy-reduction performance, the achieved reduction in received energy at the microphones of the VDA is measured. This is done by comparing the energy received when the modified signal $s_L^{(l)}$ is playing (LSp turned on) with respect to when the reference signal $s_{\text{ref}}^{(l)}$ is playing (LSp turned off). The reference playback signal is white Gaussian noise. The results are shown in Fig. 3b. In Fig. 3a it can be seen that in all scenarios the objective audio quality at the listener location reduces only slightly

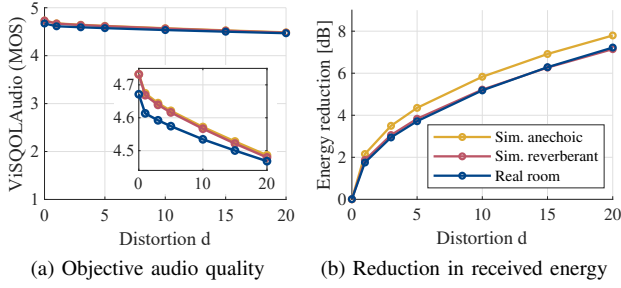


Fig. 3. The results of the objective speech quality metric (including zoom-in) at the validation points given as mean-opinion-score (MOS) (a) and the reduction in received energy at the microphones (b) as function of distortion parameter d . In (a) the presented results are averages over the different test signals and validation points. In (b) the results of the simulations are averaged over the 100 runs and the eight microphones of the array. The results of the real-world scenario are averaged over the eight microphones of the array.

as the allowed distortion increases: for all experiments the MOS score remains between ‘excellent’ (5) and ‘good’ (4). To put this in context, for our chosen maximum distortion $d = 20$, an acoustic energy reduction of minimum 7 dB (which is about a quarter of the original energy) is achieved (see Fig. 3b). And even in this case the MOS value remains high at about 4.4. Note that in both experiments depicted in Fig. 3 the performance of the algorithm in the real-world scenario is not far from the performance in simulations. Considering that in a real implementation there are several sources of inaccuracy and aspects that are not modelled these results suggest robustness of the proposed algorithm. Moreover the curve’s monotonic decrease in Fig. 3a and Fig. 3b suggest that a trade-off between energy-reduction and perceived quality can be predictably and reliably achieved using parameter d .

C. Influence of loudspeaker spotforming on ASR performance

In the previous section it is shown that the LSp can achieve a significant acoustic energy reduction around the VDA while keeping a good objective audio quality at the listener position. In this section we test our main hypothesis, namely how the addition of our novel LSp affects the ASR performance of voice driven applications. A distortion value $d = 5$ is used. This is determined empirically: by listening to the output of the LSp ourselves, we found that up to $d = 5$ the distortion remained mostly unnoticeable.

We evaluate the ASR performance via the word error rate (WER) and the word information lost (WIL) [38]. WIL ranges from 0 to 1 (lower is better) and is included because the voiced test signals introduce a large number of insertions, biasing the WER results. The results are reported for different signal-to-interferer ratios (SIRs), where the “signal” is the user voice command and the “interferer” is the combined (summed) loudspeaker signals at the reference microphone of the VDA. The ASR is Whisper Medium [39]. As user voice commands we use excerpts from the LibriSpeech database [40] with a length of at most 6 seconds. For the interferer signals we use the test signals described in Sec. IV-B. In total 2.1 hours playback runtime is made per SIR using different combinations of voice commands and interferer signals. The total time of voice commands is 0.5 hours.

Current VDAs include microphone arrays to improve the audio quality [41]. This is done using microphone beamforming algorithms. The LSp algorithm is thus evaluated by:

- 1) Tests excluding microphone beamforming. In practice, this would mean no microphone array is present in the VDA. Therefore we only use the signal from the nearest microphone (NM) to the user.
- 2) Tests including microphone beamforming. We select two beamformers: (a) MVDR beamforming with oracle (perfect) voice activity detection [7], [16] and (b) the microphone

spotformer (MS) proposed in [16]. MVDR is a classic algorithm used in many applications, therefore worth including and the LSp is inspired on the class of microphone spotformers. It is therefore interesting to test these algorithms working together. The results for the simulated reverberant environment and the real-world experiments are shown in Fig. 4.

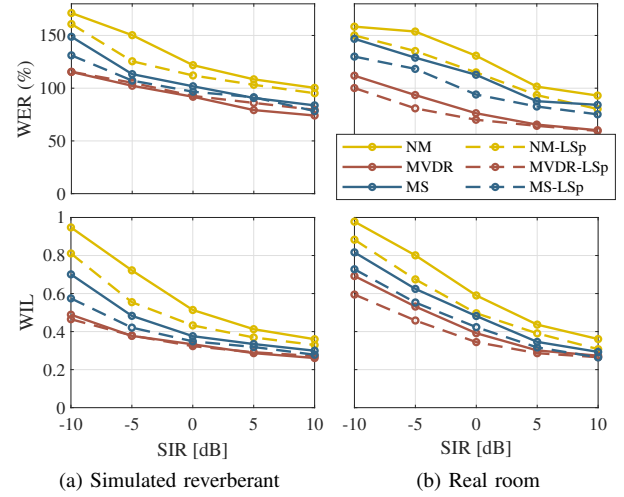


Fig. 4. The results for word error rate (WER, upper row) and word information lost (WIL, lower row) as function of SIR in simulated reverberant conditions (a) and in the real room (b), where lower is better, for the three microphone beamformers (indicated by the different colours) and with (dashed line) and without (solid line) loudspeaker spotformer (LSp). The shown results are the average results over the different test signals and voice commands. The high WER is due to the interfering speech signals, outliers were removed.

Fig. 4 shows that the average WER is predominantly more than 80% in all experiments. This highlights the difficulty of speech recognition when there are interfering speech signals at high volumes, such as when pausing a movie with dialogue using a voice command. Furthermore, the results show that using the LSp consistently improves the ASR performance on average, particularly in low SIR conditions. This confirms our main hypothesis; namely that placing the VDA in the low-acoustic-energy region created by the LSp improves ASR performance. Importantly this also holds in the real room, suggesting the robustness and applicability of the proposed algorithm.

V. CONCLUSION

We proposed a robust loudspeaker spotformer to improve speech recognition performance of voice driven applications (VDAs) in noisy backgrounds, e.g. when music is playing loudly. The proposed algorithm creates a low-acoustic-energy region around the VDA while limiting the distortion introduced to the user of the VDA. Our main hypothesis was that if the VDA is placed in the low-acoustic-energy region the automatic speech recognition (ASR) performance increases. Our experiments showed that the low-acoustic-energy region can be created while keeping the objective audio quality high. Furthermore, the algorithm provides a parametric trade-off between acoustic energy reduction and reproduction quality. We confirmed our main hypothesis, namely the loudspeaker spotformer can improve ASR performance. We verified this in a real room, suggesting the robustness of our algorithm. A limitation of this work is the computational complexity involved in solving optimisation problem (12). Therefore an efficient or analytic solution for (12) is a topic for future work. Additionally, the perceived audio quality at the user location should be confirmed through subjective tests. Lastly, other aspects of reproduction quality such as preservation of interaural-time and -level differences should be tested.

REFERENCES

- [1] B. King, I. Chen, Y. Vaizman, Y. Liu, R. Maas, S. H. K. Parthasarathi and B. Hoffmeister, "Robust Speech Recognition Via Anchor Word Representations," *Proc. Interspeech 2017*, Stockholm, Sweden, 2017, pp. 2471-2575.
- [2] V. A. Vu and M. Akselrod, "An experiment of dual-LTE MPTCP with In-Car Voice Assistant," *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, Helsinki, Finland, 2021, pp. 1-5.
- [3] B. Minder, P. Wolf, M. Baldauf and S. Verma, "Voice assistants in private households: a conceptual framework for future research in an interdisciplinary field," *Humanit Soc Sci Commun* 10, 173, 2023.
- [4] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix and T. Nakatani, "Far-Field Automatic Speech Recognition," in *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124-148, Feb. 2021.
- [5] Y. Gong, S. Khurana, L. Karlinsky and J. Glass, "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers," *Proc. Interspeech 2023*, 2023, pp. 2798-2802.
- [6] A. Rouditchenko et al., "Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation," arXiv:2406.10082 [eess.AS], June 2024.
- [7] S. Gannot, E. Vincent, S. Markovich-Golan and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692-730, April 2017.
- [8] H. A. Kassir, Z. D. Zaharis, P. I. Lazaridis, N. V. Kantartzis, T. V. Yioultsis and T. D. Xenos, "A Review of the State of the Art and Future Challenges of Deep Learning-Based Beamforming," in *IEEE Access*, vol. 10, pp. 80869-80882, 2022.
- [9] C. Quan and X. Li, "SpatialNet: Extensively Learning Spatial Information for Multichannel Joint Speech Separation, Denoising and Dereverberation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1310-1323, 2024.
- [10] H. Chen et al., "The First Multimodal Information Based Speech Processing (Misp) Challenge: Data, Tasks, Baselines And Results," *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 9266-9270.
- [11] M.E. Sadeghi, H. Sheikhzadeh and M. J. Emadi, "A proposed method to improve the WER of an ASR system in the noisy reverberant room," in *Journal of the Franklin Institute*, vol. 361 (1), pp. 99-109, 2024.
- [12] C. Manzanillo, R. Chettiar, R. Soroushmojdehi, L. Ying, J. Dong and M. Mohsenvand, "Soft Speech, Loud World: Bone Conduction Microphones Enhance Voice Assistant Interaction," *2024 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, 2024, pp. 1-5.
- [13] L. Fan, L. Xie, X. Lu, Y. Li, C. Wang and S. Lu, "mmMIC: Multi-modal Speech Recognition based on mmWave Radar," *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, New York City, NY, USA, 2023, pp. 1-10.
- [14] S. Miyabe, Y. Hinamoto, H. Saruwatari, K. Shikano and Y. Tatekura, "Interface for Barge-in Free Spoken Dialogue System Based on Sound Field Reproduction and Microphone Array," *EURASIP J. Adv. Signal Process.* 2007, 057470.
- [15] M. Taseska and E. A. P. Habets, "Spotforming: Spatial Filtering With Distributed Arrays for Position-Selective Sound Acquisition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1291-1304, July 2016.
- [16] J. Martinez, N. Gaubitch and W. B. Kleijn, "A robust region-based near-field beamformer," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 2494-2498.
- [17] T. Lee, J. K. Nielsen and M. G. Christensen, "Signal-Adaptive and Perceptually Optimized Sound Zones With Variable Span Trade-Off Filters," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2412-2426, 2020.
- [18] N. de Koeijer, M. B. Møller, J. Martinez, P. Martínez-Nuevo and R. C. Hendriks, "Block-Based Perceptually Adaptive Sound Zones with Reproduction Error Constraints," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- [19] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen and S.H. Jensen, "A Perceptual Model for Sinusoidal Audio Coding Based on Spectral Integration". *EURASIP J. Adv. Signal Process.* 2005, 317529 (2005).
- [20] D. de Groot, 2025, "Code related to "Loudspeaker Beamforming to Enhance Speech Recognition Performance of Voice Driven Applications"", 4TU.ResearchData. [Online]. Available: <https://doi.org/10.4121/36b9065e-278e-40ee-b359-6cd734561f86>.
- [21] R. Heusdens and N. Gaubitch, "Time-delay estimation for TOA-based localization of multiple sensors," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 609-613.
- [22] J. Ahrens, *Analytic Methods of Sound Field Synthesis*, Springer, Berlin, Ger., Jan. 2012.
- [23] M.H. Hayes. Discrete-Time Random Processes. In: *Statistical digital signal processing and modeling*. John Wiley & Sons, 1996.
- [24] R. Horn and C. Johnson, *Matrix Analysis (2nd Ed.)*. Cambridge: Cambridge University Press, 2012.
- [25] M. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, June 2001.
- [26] N. L. Johnson, S. Kotz and N. Balakrishnan, *Continuous Univariate Distributions, Volume 1 (2nd Ed.)*, Wiley, 1994.
- [27] S. Boyd and L. Vandenberghe, *Convex Optimisation*, Cambridge University Press, 2009.
- [28] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.2. <https://cvxr.com/cvx>, 2014.
- [29] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach (1st Ed.)*, CRC Press, 2003.
- [30] A. Jeannerot, N. de Koeijer, P. Martínez-Nuevo, M. B. Møller, J. Dyreby and P. Prandoni, "Increasing Loudness in Audio Signals: A Perceptually Motivated Approach to Preserve Audio Quality," *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 1001-1005.
- [31] J.O. Smith, *Spectral Audio Signal Processing*. <http://ccrma.stanford.edu/~jos/sasp/>, online book, 2011 edition, accessed Apr. 2024.
- [32] W. H. Press, B. P. Flannery, A. S. Teukolsky, W. T. Vetterling, *Numerical Recipes in C : The Art of Scientific Computing*, Cambridge University Press, Oct. 1992.
- [33] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943-950, Apr. 1979.
- [34] E. A. P. Habets, "Room impulse response generator", Sept. 2010. GitHub Repository, <https://github.com/ehabets/RIR-Generator>, accessed May 2023.
- [35] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram and N. Harte. "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," in *J. Acoust. Soc. Am.* vol. 137, no. 6, pp. EL449-EL455, 2015.
- [36] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman and A. Hines, "ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric," *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, Athlone, Ireland, 2020, pp. 1-6.
- [37] Methods for subjective determination of transmission quality, Rec. ITU-T P.800, International Telecommunications Union, Geneva, Switzerland, 1996.
- [38] A. C. Morris, V. Maier and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," *Proc. Interspeech 2004*, 2004, pp. 2765-2768.
- [39] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv:2212.04356 [eess.AS], Dec. 2022.
- [40] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 5206-5210.
- [41] X. Ji, G. Zhang, X. Li, G. Qu, X. Cheng and W. Xu, "Detecting Inaudible Voice Commands via Acoustic Attenuation by Multi-channel Microphones," in *IEEE Transactions on Dependable and Secure Computing*, 2024.