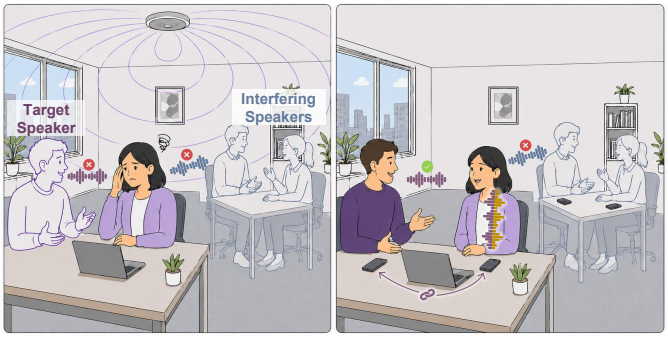
Focus on What You Want: Individual Intelligibility Regulation via Selective Sound Masking

***Abstract*—Conversations in shared spaces often cause inter-group interference, which distracts the participants. Sound masking, a widely deployed technique, can reduce such interference by lowering speech intelligibility. Existing systems, however, are non-selective: a fixed masking signal affects all speakers alike, degrading both interfering and desired speech. We present AMasker, a system that makes sound masking selective. AMasker exploits the fact that a masking sound tailored to one speaker is naturally less effective against another. Combining speaker-aware signal design and speech separation technologies, AMasker generates personalized masking signals that selectively reduce unwanted speech intelligibility. We implement a prototype on commodity devices and evaluate it across diverse conditions. Results show that, in both objective and subjective evaluations, AMasker consistently outperforms baselines in reducing interfering intelligibility while improving target clarity in real time.**

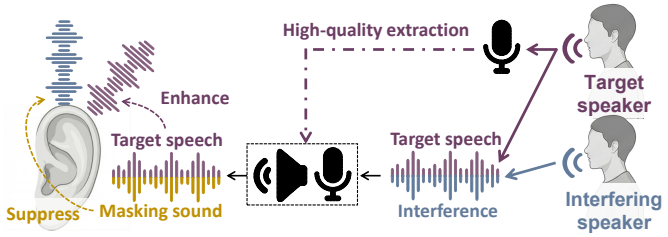
I. INTRODUCTION

Verbal conversations frequently occur in places such as open-plan offices, cafeterias, seminar rooms, and even vehicles. Due to space limitations, conversation groups may coexist in a single space, leading adjacent groups to interfere with each other. Such interference pervades the shared environment, diverting participants' attention and impeding their ability to concentrate on their own conversations. Indeed, background speech has been a primary source of dissatisfaction for workers, which negatively affects employee well-being, job satisfaction, and work performance [1], [2].

A key observation underlying speech interference is that distraction depends more on speech intelligibility than on loudness [1], [3]. In other words, a background conversation becomes distracting primarily when its content can be understood, rather than simply because it is loud. Sound masking is an established technique for mitigating distraction in shared workspaces [4], [5]. It builds on the masking effect [6]: human auditory perception of a speech signal is reduced by a louder sound played in close temporal proximity. For example, a whisper that distracts in a quiet library becomes imperceptible under heavy rainfall. Sound masking thus reduces distraction by suppressing intelligibility, even though it raises the overall sound level. However, existing sound masking systems [7], [8] lack selection capability. They continuously emit fixed signals such as pink or white noise [2], which lower the intelligibility of all speech indiscriminately (Fig. 1 (a)), even the target speech. This contradicts the requirement in conversation scenarios, where we need to keep the intelligibility of the target speech that a listener is trying to focus on. Thus, we need a **selective sound masking** method to mask only the undesired speech with minimal negative impact on the target speech.



(a) Non-selective sound masking (left) vs. selective masking (right).



(b) The basic idea.

Fig. 1: Illustration of AMasker: individual intelligibility regulation with general audio devices.

We propose AMasker, a novel system that generates highly customized masking sounds to mask only the undesired speech. Our key insight is that, due to the difference in speakers' voice characteristics, speech content, and speaking rhythm, the target and the interfering speech diverge substantially in spectrum [9], [10]. Given that sound masking depends critically on spectral alignment between the masking and the masked sound [6], [11], *a masking sound tailored to one speaker is naturally less effective against another*. Thus, if we can accurately discriminate the two sounds and generate the masking sound specifically for the interfering speech, only while it is present, the masking sound remains effective against the interference while being less harmful to the target speech.

AMasker can run on commodity devices with speakers and microphones (e.g., users' smartphones). Fig. 1 (b) illustrates the basic idea of AMasker, which leverages collaboration among users' devices to create a mutually masked environment for conversation groups. Specifically, each device separates the speech it is best placed to capture. After extracting the target and the interfering speech separately, AMasker emits a signal that is a superposition of the separated target speech for target enhancement and a carefully designed masking sound for interference suppression. Three challenges arise in

implementing AMasker.

Challenge 1: *Efficient masking sound generation without intelligibility leakage.*

An efficient masking sound should suppress the interference's intelligibility with minimal added energy, which favors tight spectral alignment. However, short-frame real-time masking sound generation with tight spectral alignment will introduce intelligibility leakage from the masked speech to the masking sound, making the masking sound a distraction itself.

We solve this problem with a spectral geometric reshaping method. It generates a masking sound that preserves the coarse-grained spectral trends of the interference for masking efficiency, while deliberately smearing the intelligibility-critical fine-grained details to suppress the leakage.

Challenge 2: *Target speaker identification in multi-speaker conversations.*

Because the masking sound inevitably affects target speech that overlaps the interfering speech, AMasker also extracts and replays the target speech. This requires the target speaker's voiceprint. However, although users' voiceprints can be collected in advance, the active speaker switches frequently in a multi-speaker conversation. It is difficult to identify the target speaker in time and fetch the correct voiceprint.

To address this problem, we propose a collaboration-based speaker identification method, where distributed devices jointly determine the active speaker in real time by combining DNN-based identification with spatial cues (i.e., TDoA).

Challenge 3: *Interfering speech separation without the interferers' voiceprints.*

Generating the masking sound requires extracting the interfering speech. However, the interferers' voiceprints are out of reach: each user's voiceprint is shared only within their own conversation group for privacy, and interferers may not even be registered users. Moreover, the interfering speech must be told apart from non-intelligible noise, which needs no masking.

We address this with a dual-decoder speech separation architecture that outputs the target and the interfering speech simultaneously based on the target speaker's voiceprint alone, treating all interfering speech as a whole—we only need to mask the intelligible sound beyond the target. Non-intelligible noise is injected during training so that the model can learn to reconstruct only the intelligible interference.

We prototype AMasker and conduct comprehensive experiments in real scenarios, including crowded public places and an in-vehicle environment. The results show that AMasker reduces the target word error rate by more than 4× relative to baselines without target enhancement capabilities, while reducing the intelligibility of interfering speech by 75.8% relative to baselines without interference suppression capabilities. Subjective evaluations further demonstrate that AMasker effectively helps users direct their attention to the target speaker, outperforming all baselines. It also meets real-time requirements, with a processing delay below half of the latency budget. We summarize our contributions as follows:

- We propose AMasker, the first selective sound masking system that suppresses the intelligibility of the interfering speech while enhancing the target speech.
- We design a spectral geometric reshaping method that eliminates intelligibility leakage in real-time masking sound generation. We further design a collaboration-based speaker identification method and a dual-decoder separation architecture that extract both the target and the interfering speech with only the target speaker's voiceprint.
- We present a working prototype of AMasker and validate its effectiveness in extensive real-world scenarios.

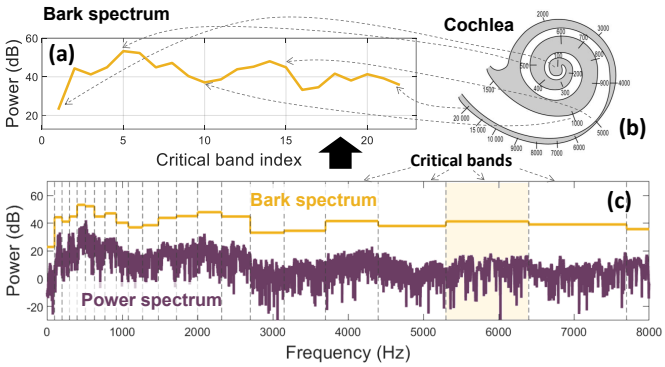


Fig. 2: Power spectrum and Bark spectrum.

II. RELATED WORK

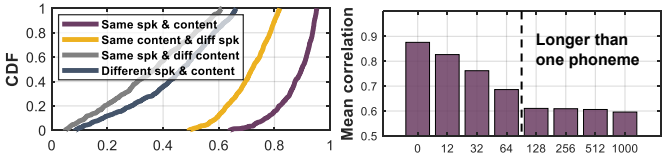
Headphone-based selective hearing. Existing works leverage speech separation technologies to make a listener wearing headphones hear only selected speakers—a chosen one [12] or those within a "sound bubble" [13]—by suppressing all sounds with ANC-enabled headphones and replay only the selected sound. These methods only serve the specialized hearables; AMasker instead serves the devices users already own.

Sound masking systems. Sound masking has attracted considerable attention and recognition in academia [11], [1] and has been successfully deployed in renowned companies [14], receiving highly positive evaluations for its distraction-suppression performance. However, most existing approaches rely on stationary masking signals—ranging from conventional broadband noise to natural sounds [15], [16]—which indiscriminately mask all speech. In contrast, AMasker introduces advanced speech separation and adaptive masking-sound generation, enabling selective sound masking and striking an optimal balance between interference suppression and target-speech enhancement.

III. BACKGROUND AND INTUITION

A. Masking effect and Bark spectrum

The masking effect originates from the cochlea's frequency-resolving mechanism. When two sounds are close in frequency, they excite overlapping regions on the basilar membrane. The louder sound (the "masker") suppresses the neural response to the weaker sound (the "maskee"). Moreover, the auditory system acts as a set of bandpass filters whose bandwidth—the *critical band*—reflects the cochlea's frequency resolution [17]. Sounds within the same critical band are processed as a single entity, making them hard to resolve individually. Harnessing this principle, sound masking



(a) Bark variation across different speakers and contents

(b) Bark variation within the speech of a single speaker

Fig. 3: Bark spectrum analysis across different conditions.

sends a controlled masking sound[1] to reduce the intelligibility of interfering speech.

Masking sound generation and evaluation are typically performed based on Bark spectrum analysis [18], which divides the audible frequency range into 24 critical bands, matching auditory perception more closely than other representations (e.g., the Mel spectrum) [19]. Because components inside one critical band are not resolved individually, the perceptually meaningful quantity there is their total energy rather than the individual spectral bins. The Bark spectrum therefore represents a signal by the power summed within each band i: $B(i) = \sum_{k=bl(i)}^{bh(i)} P[k]$, where $bl(i)$ and $bh(i)$ are the band's frequency boundaries and $P[k]$ is the power of the k-th frequency bin. Fig. 2 shows an example. The bandwidth of critical bands increases with frequency because human auditory resolution is coarser at higher frequencies.

B. Feasibility study

Since the Bark spectrum reflects how the human auditory system perceives sound across critical bands, we examine this perception from the Bark perspective across speakers/content and over time in this section. We demonstrate key properties of the speech Bark spectrum that validate the feasibility of selective sound masking.

Property 1: *Bark-spectrum characteristics of speech vary substantially across speakers and phonetic contents.*

We verify this property via a comparative analysis on the TIMIT dataset [20] (which provides detailed annotations of speakers and speech content). We compute Pearson correlations between Bark spectra of 500 pairs of 12-ms frames under four conditions: i) *same speaker & content*, pairing temporally adjacent frames within the same phoneme from the same speaker; ii) *same content & different speakers*, pairing the same phoneme across different speakers; iii) *same speaker & different content*, pairing different phonemes from the same speaker; iv) *different speakers & different content*, pairing different phonemes from different speakers.

Fig. 3 (a) shows the CDF of the Pearson correlation coefficient between Bark spectrum pairs obtained under different conditions. As shown, the Bark spectrum of speech varies dramatically across speakers and speech content.

Property 2: *Bark-spectrum characteristics of speech exhibit rapid temporal variation.*

We examined this by selecting 100 speech segments (10s each), dividing them into 12-ms frames, and computing Pearson correlations between Bark spectra of frames separated by varying time intervals.

Fig. 3 (b) shows how this correlation changes with the time interval. The Bark spectra of two frames decorrelate fast even within a single phoneme ($\approx$ 100 ms [21]).

Insights. Our analysis yields key conclusions for system design: i) human speech contains sufficient spectral diversity for auditory distinction; ii) masking sounds must adapt in real time to the time-varying spectral characteristics of interfering speech. These make it feasible to generate masking sound that selectively masks a specific speech.

IV. SYSTEM OVERVIEW

Fig. 4 summarizes AMasker's system architecture, which consists of three core components. As a prerequisite, user embeddings, which contain their voiceprints, are collected during daily smartphone interactions (e.g., phone calls) and exchanged among participants before a conversation begins, forming a shared embedding table. During the conversation, the speaker identification module determines the current speaker and fetches the matching embedding, based on which the speech separation module extracts the target and interfering speech from the mixture. The masking sound generation module then produces the masking sound, and the device emits it together with the replayed target speech, suppressing the interference while enhancing the target speech.

Next, we describe how to generate the effective masking sound from the extracted speech (Sec. V), followed by the speech separation method supporting this process (Sec. VI).

V. MASKING SOUND GENERATION

We first introduce important metrics. The Speech Intelligibility Index (SII) is typically used to evaluate the performance of sound masking. It is an index ranging from 0 to 1 that quantifies how much of an interfering sound S_I remains intelligible in the presence of a masking sound M. Psychoacoustic studies show that *an SII below 0.2 indicates that the interfering speech is effectively masked* [22]. Following ANSI S3.5-1997 [23], the SII weights each critical band by its contribution to intelligibility and by how audible the interfering speech remains there: $\sum_i I_i \cdot \min(1, \max(0, (R_i + 15)/30))$, where I_i is the band importance function ($\sum_i I_i = 1$) and R_i is the SNR of S_I against M within critical band i. A band contributes nothing when $R_i \leq -15$ dB and its full weight I_i when $R_i \geq 15$ dB, varying linearly in between.

A louder masking sound always lowers SII, but sound levels above 85 dBA can damage the human auditory system [24]. Moreover, since the target and the interfering speech may overlap, without loudness control, the masking sound will substantially reduce the target speech intelligibility. So, we should also minimize the masker-to-speech ratio (MSR): $10\log_{10}(Sum(P(M))/Sum(P(S_I)))$, where $P(M)$ and $P(S_I)$ are the power spectra of masking sound and interfering speech, respectively. An effective masking sound should suppress interfering speech intelligibility (i.e., SII $<$ 0.2) with minimal added loudness (i.e., MSR $<$ 20 dB, since speech is typically 55-65 dBA [25]).

[1]We use "masking sound" to represent "masker" for better understanding.

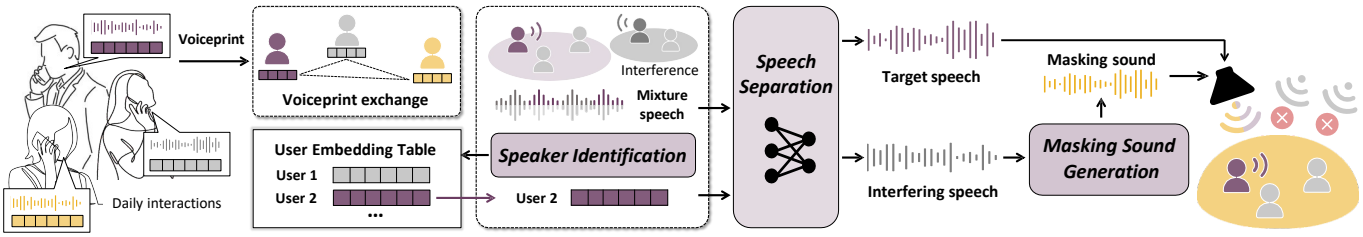


Fig. 4: Overview of AMasker.

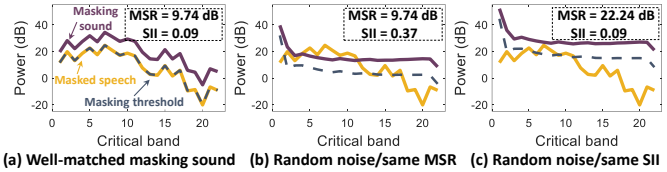


Fig. 5: Impact of masking threshold's alignment.

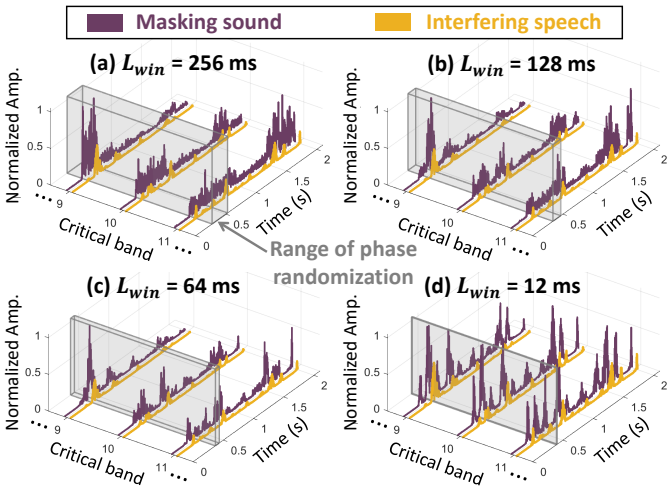


Fig. 6: Comparison of spectro-temporal structures across different processing window lengths.

A. The efficiency–leakage dilemma

The energy-efficient masking sound. An interfering speech is well masked when its Bark spectrum is under the masking threshold of a masking sound [17], [18], which is a Bark-scaled energy level across critical bands below which the interfering speech becomes unperceived. Within a critical band, once the threshold exceeds the interferer's Bark energy, raising the threshold further yields no additional suppression and only adds loudness. The energy-efficient masking sound is therefore the one whose masking threshold just covers the interfering speech's Bark spectrum in every critical band with no excess margin. Fig. 5 illustrates this: such a masking sound suppresses the interference to SII = 0.09 at an MSR of only 9.74 dB, whereas a random noise at the same loudness attains only SII = 0.37, and reaching SII = 0.09 costs it an MSR of 22.24 dB. We thus obtain this masking sound M^* by treating the interfering speech's Bark spectrum as the target masking threshold and numerically inverting the standard masking-sound-spectrum-to-threshold computation [18], [26], which yields M^*'s Bark spectrum B_M^*.

TABLE I: WER of ASR transcribing the masking sound itself.

Type	M^*	M^*	M^*	M^*	M^*	$M^{\#}$
L_{win} **(ms)**	12	32	64	128	256	12
WER	0.041	0.043	0.136	0.844	0.959	0.979

The intelligibility leakage problem. Following the interference this closely causes *intelligibility leakage*: the spectral envelopes of the two sounds become so similar that the masking sound *emulates* the interfering content and turns intelligible itself. An effective method to address this problem is to randomize the phase of M^*. *Speech content is encoded by spectro-temporal cues across narrow frequency bands* [27], [28], carried jointly by two components: i) the *temporal envelope* in each band, whose modulations govern articulatory patterns such as syllable rhythm and segmentation; and ii) the *spectral envelope*, which specifies the energy distribution across bands and characterizes phonetic content, including vowel and consonant identity. So, although M^* is highly correlated with interfering speech in the frequency domain, the phase randomization severely scrambles the time-domain signal within each frequency band, smearing the temporal envelope and destroying intelligibility. Fig. 6 (a) confirms this: with the phase scrambled within 256-ms windows, the temporal envelopes of M^* in the key critical bands (920-1480 Hz) are severely blurred relative to the interfering speech.

Tension under the real-time constraint. Phase scrambling, however, works only when the processing window L_{win} is long enough. A short L_{win} limits the scrambling range and leaves intact the rapid rhythmic changes (e.g., the sharp peaks and valleys in temporal envelopes) that carry the intelligibility-critical cues. Figs. 6 (b)-(d) show temporal envelopes of the energy-efficient masking sound M^* generated by different L_{win}. As expected, M^* generated with a shorter L_{win} is more correlated with the interfering speech. Since the Word Error Rate (WER) of speech achieved by an Automatic Speech Recognition (ASR) model is widely accepted to assess the intelligibility [29], we further use the SOTA model Whisper-Large-V3 [30] to evaluate the intelligibility of M^* generated with different L_{win}. The results in Table I indicate that L_{win} must reach 128 ms to keep M^* unintelligible (i.e., WER $>$ 0.5 [31]). Real time demands the opposite. A longer window raises latency and desynchronizes the masking sound from the interfering speech. We apply varying delays to masking sounds generated for 40 LibriSpeech speakers, and Fig. 7 shows the delay must not exceed 32 ms to hold SII below 0.2. Based on results in Sec. VII-F, we set the processing window L_{win} to

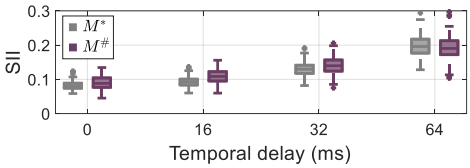


Fig. 7: Impact of temporal delay on masking.

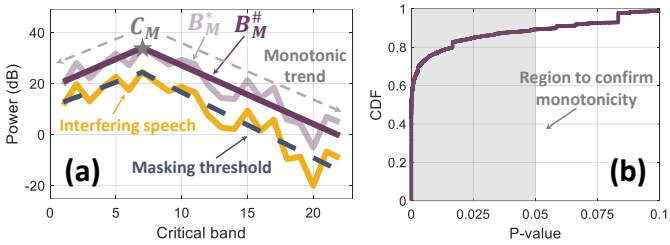


Fig. 8: (a) Illustration of the proposed Bark reshaping method; (b) Results of monotonicity test.

12 ms, yielding an average processing latency of 9.52 ms.

B. Spectral geometric reshaping

Instead of scrambling the time-domain signal, we fine-tune the Bark spectrum B_M^* of the energy-efficient masking sound, reducing its spectral similarity with the interference while retaining its masking effect. Casting this as a multi-objective optimization is impractical: iterative solvers (e.g., gradient descent) are far too slow for real time, and no metric can directly measure the intelligibility of a ms-level masking sound to guide them.

Insight. A geometric property of B_M^* lets us bypass complex optimization methods. Although B_M^* varies with the interfering speech's content, *its shape exhibits a special "pyramid-like" pattern*. As Fig. 8 (a) shows, B_M^* splits at the critical band with maximum energy, denoted as C_M, with the Bark bins on each side following a monotonic trend. A Spearman's rho test [32] over the Bark spectra of masking sounds generated based on the LibriSpeech dataset confirms this. Fig. 8 (b) shows that nearly 90% of samples yield $p < 0.05$ for monotonicity on both sides of C_M. Two linear fittings therefore suffice to smear the intelligibility-critical fine-grained details of the spectral envelope while maintaining the coarse-grained spectral trends that masking efficiency relies on.

Bark reshaping. Taking C_M as the boundary, we fit a line to the Bark bins on each side by the least squares method. The two lines, intersecting at C_M, form the new Bark spectrum $B_M^\#$, as shown in Fig. 8 (a). We further scale $B_M^\#$ to match the loudness of B_M^*, and use $B_M^\#$ to reconstruct the new power spectrum $P^\#(M)$ and the masking sound $M^\#$.

Result. Fig. 7 also reports $M^\#$ under the same delay setting. Its SII stays comparable to that of M^* and below the 0.2 criterion within the 32-ms tolerance, confirming that the reshaped masking sound retains the masking effect. Table I further shows $M^\#$ to be more unintelligible than M^* generated with a 256-ms phase randomization window. $M^\#$ thus attains both objectives at a 12-ms window.

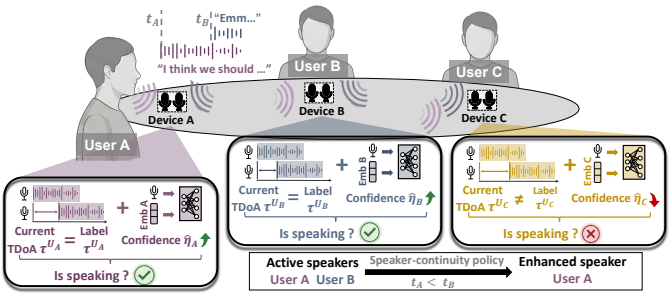


Fig. 9: The collaborative identification scheme.

VI. SPEECH SEPARATION

Real-time speech separation that accurately extracts both target and interfering components from a mixture is essential for selective sound masking. This hinges on two key steps: first, obtaining a reliable target speaker embedding, and second, performing online separation of target and non-target speech conditioned on that embedding [33].

Embedding acquisition. AMasker employs a lightweight embedding model suitable for resource-constrained devices [12], taking a 5-second voice sample as input. User recordings are obtained in two ways: (i) users proactively provide a sample, and (ii) with proper permissions, AMasker runs as a background service on personal devices, collecting high-quality speech during everyday interactions such as phone calls. When a conversation begins, participants' devices exchange embeddings to form a shared table, ensuring reliable target-speaker identification even in multi-party settings.

Spatially distributed speech separation. To improve separation performance, target speech is extracted on the speaker's device, which is physically closer to the speaker and thus captures the target speech with higher SNR and direct-to-reverberant ratio [34]. The separated target speech is then relayed to all listeners via the local network. In contrast, the latency-critical masking path, from interfering-speech separation to masking-sound generation, runs entirely on each listener's device, incurring no network latency. The separated interfering speech also better matches what the listener actually hears, enabling more precise masking.

In this section, we first introduce the speaker identification method to select the correct embedding as a prerequisite for speech separation, then present a dual-decoder architecture that accurately extracts both target and interfering speech.

A. Robust speaker identification

A straightforward speaker identification approach would require each device in the group to individually verify all group members' identities to locate the active speaker, incurring significant computational overhead and latency. To avoid this, as shown in Fig. 9, AMasker adopts a *collaborative identification scheme*, where each device only determines whether its user is speaking, and broadcasts a status message $\gamma_{U_i} \in \{0, 1\}$ (0 for silence, 1 for speaking). Upon receiving messages from other in-group devices, each device asynchronously updates the speaker embedding used by subsequent processing.

Speaker identification with hybrid cues. Speaker identification must operate on sub-second speech windows to improve user experience, yet such brief segments contain limited speaker-specific information for reliable identification. AMasker addresses this challenge with a hybrid identification design fusing two complementary cues: i) speaker-specific voice characteristics via a DNN-based identification module, which takes a user's embedding and a short audio segment to determine whether the segment contains that user's speech; ii) spatial evidence from time-difference-of-arrival (TDoA) via the ubiquitous dual-microphone setup. These two modalities are naturally complementary: the DNN module captures voice characteristics that remain effective under mobility, whereas TDoA offers stable spatial evidence when user–device positions remain approximately constant. To fully exploit this synergy, we propose a quality-adaptive fusion method. For each detected speech window, AMasker computes the fused speaking score for its user U_i as:

$$S_i(t) = \alpha(t)\hat{\eta}_i(t) + (1-\alpha(t))\phi\left(\left|\tau_i(t) - \tau_L^{U_i}\right|\right) \quad (1)$$

where $\hat{\eta}_i(t) \in [0,1]$ is the model's confidence, $\tau_i(t)$ is the current TDoA estimate, and $\tau_L^{U_i}$ is the user's TDoA label, initialized from high-confidence windows at the conversation's start and dynamically updated upon detecting spatial offsets at runtime. ϕ is the TDoA matching score, positive when the measured TDoA aligns with the stored label and negative otherwise. The weight $\alpha(t)$ adapts to TDoA quality, jointly evaluated from the cross-correlation peak sharpness and temporal stability of TDoA estimates across recent windows. High-quality TDoA measurements yield a small $\alpha(t)$, letting spatial cues dominate the speaking score, whereas the DNN model assumes primary responsibility under mobility or low-SNR conditions. For robustness, γ_{U_i} is set to 1 only when $S_i(t)$ exceeds a threshold for three consecutive windows. Empirically, the window length and hop size are 500 ms and 25 ms, respectively.

Handling overlapped in-group speakers. In natural conversations, speaker overlap may occasionally occur, causing multiple devices to broadcast $\gamma_{U_i} = 1$. Since AMasker replays extracted speech to listeners, replaying multiple overlapping speeches contradicts the goal of directing attention to a target speaker. AMasker therefore adopts a *speaker-continuity policy*: upon overlap, the system retains the turn-initiating speaker as the target and treats subsequent overlapping speech as interference. This aligns with the well-established turn-taking structure of human conversation, where most overlaps are transient (lasting only a few hundred milliseconds [35]) and do not immediately alter the current dominant speaker [36].

B. Joint target–interference separation

The separation model has to deliver the interfering speech for masking sound generation, without the interferers' voiceprints and without mistaking non-intelligible noise for interference. Two observations make this attainable with the target embedding alone. First, masking needs only the aggregate interference, so separation reduces to a binary split—the target versus all other intelligible sound—whose boundary the target embedding suffices to delineate. Second, since intelligible speech exhibits distinctive spectro-temporal modulation cues, learned models can reliably distinguish it from non-speech sounds [37].

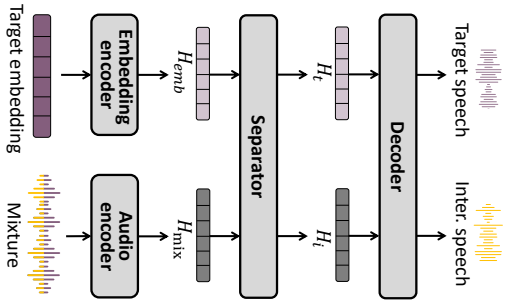


Fig. 10: The dual-decoder architecture.

Dual-decoder architecture. As shown in Fig. 10, we instantiate this binary split with the encoder-separator-decoder framework [38], extended with two output branches. The audio encoder and the embedding encoder first transform the mixture speech and the target embedding into latent representations as H_{mix} and H_{emb}. Then, the separator extracts the estimated target and interfering representations, H_t and H_i, from H_{mix}, conditioned on H_{emb}. The decoder further reconstructs the target and interfering speech based on H_t and H_i. Jointly emitting both streams disentangles the interference from the target speaker's representation space, preventing interfering patterns from leaking into the target features. We adopt SNR loss [39] alongside a standard contrastive term for training:

$$\mathcal{L} = -\beta_1 \,\mathrm{SNR}(s_t, \hat{s}_t) - \beta_2 \,\mathrm{SNR}(s_I, \hat{s}_I) + \beta_3 \,\mathcal{L}_{cl} \quad (2)$$

where $\hat{s}_t$ and s_t are the estimated target speech and its ground truth. $\hat{s}_I$ and s_I are the estimated interfering speech and its ground truth. $\mathcal{L}_{cl}$ is a contrastive loss adopted from prior work [40]: it pulls the latent target features toward their ground truth while pushing them away from the interfering features, reinforcing the feature disentanglement. β_1, β_2, and β_3 are adjusted by an adaptive scaling method [41].

Data augmentation. We simulate the deployment conditions during training: speech is convolved with diverse Room Impulse Responses, and non-intelligible noise (e.g., white noise and synthetic masking sounds) is added. This augmentation steers the model to extract only intelligible components, preventing it from mistakenly treating masking sounds emitted by nearby devices as interference. This also avoids the positive feedback loop in which devices continuously raise their masking-sound levels in response to each other.

VII. EVALUATION

A. Implementation

We prototype AMasker using commodity devices (Sec. VII-B). The signal is sampled at 48 kHz for accurate TDoA estimation and downsampled to 16 kHz for other tasks. An acoustic echo canceller suppresses the self-induced masking

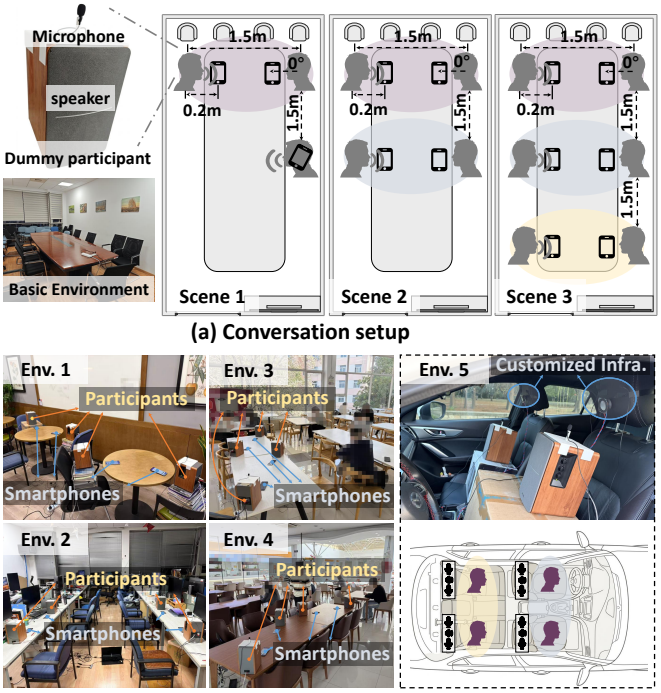


Fig. 11: Experiment settings.

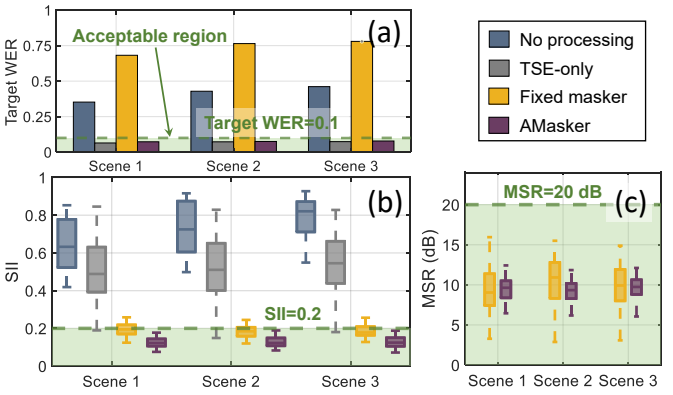


Fig. 12: Comparison with baselines in terms of (a) target enhancement; (b)∼(c) interference suppression.

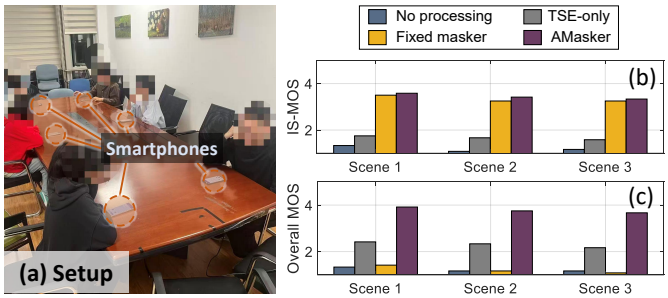


Fig. 13: Comparison with baselines in human perception.

sound in the captured signal. Audio I/O runs in AAudio MMAP mode to avoid extra buffering. The replayed target speech is amplified to maintain a 5 dB gain over the generated masking sound. We implement the collaboration network over Wi-Fi, using UDP for real-time transmission.

B. Experimental methodology

1) Experiment settings: For objective evaluation, we replace humans with dummy participants consisting of commercial speakers and microphones (see Fig. 11 (a)), enabling reproducible real-world conversations. Each dummy participant is paired with a smartphone, and replays speech calibrated to 60 dBA at 1 m with a sound pressure level meter, matching typical conversational levels [25]. As Fig. 11 (a) shows, conversations run across three scenes in a basic environment, with interference levels increasing from Scene 1 to 3. Fig. 11 (b) shows five unseen test environments. Env. 5 tests whether AMasker, being software-only, generalizes from smartphones to general audio devices: we build an audio system from commercial microphones and speakers as the vehicle's existing infrastructure. For subjective evaluation, approved by our IRB, 12 volunteers (7 males and 5 females) are invited to form groups and converse, as shown in Fig. 13 (a).

2) Conversation synthesis: We synthesized 100 random-content conversations from the LibriSpeech dataset, assigning each participant random voice characteristics. Each conversation lasts 3-5 minutes, with each speaking turn lasting 5-15 seconds. To better simulate natural speech overlap [35], we insert random 100–400 ms overlaps at turn boundaries and mid-turn intrusions of 300–500 ms, which cumulatively account for 5% of the total conversation duration.

3) Metrics: Comprehensive metrics are adopted as:

i) *Word Error Rate (WER)*: We calculate the ASR WER of the target speech with Whisper-Large-V3. WER below 0.1 is considered acceptable for human understanding [42].

ii) *Speech Intelligibility Index (SII)*: We calculate SII as introduced in Sec. V. Note that SII should stay below 0.2.

iii) *Masker-to-Speech Ratio (MSR)*. We calculate MSR as introduced in Sec. V, and it should be lower than 20 dB.

For subjective evaluation, we collect mean opinion score (MOS) from volunteers to evaluate the ability to regulate a listener's attention towards the target speaker. Volunteers rate performance by answering the following questions:

i) *Interference suppression MOS (IS-MOS):* How intelligible is the interfering speech? *1 - Always intelligible, 2 - Sometimes intelligible, 3 - Perceptible, but little intelligible, 4 - Sometimes perceptible, 5 - Not perceptible;*

ii) *Overall MOS:* If the goal is to focus on the target speaker without distraction, how was your overall experience? *1 - Bad, 2 - Poor, 3 - Fair, 4 - Good, 5 - Excellent.*

C. Overall performance

Here, we introduce an evaluation framework for intelligibility control capability and demonstrate that AMasker outperforms baselines in enhancing listeners' attention to the target speech. The baselines are:

i) *No processing*: There is no specific method to suppress interference or enhance target speech.

ii) *Target speech enhancement (TSE-only)*: We take the target speech extraction model of existing work [12] as a representative prior solution, retrained on the same dataset

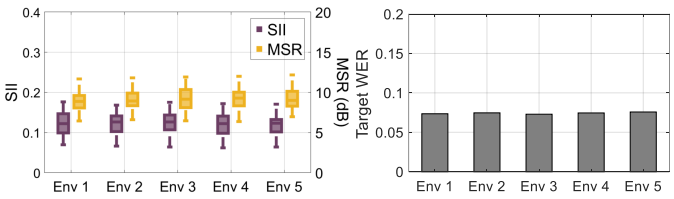


Fig. 14: Performance under different environments.

and with the same monaural input as AMasker. It replays the extracted target speech at 10 dB above the interfering speech, without interference suppression.

iii) *Fixed masker*: This method broadcasts a fixed pink noise signal whose energy keeps 10 dB higher than the environmental sounds.

Comparison of target enhancement. For each method, we test under different conversational scenes in Fig. 11 (a). Target WER results appear in Fig. 12 (a). For unprocessed speech, the target WER exceeds 0.35 across all scenarios, as both the target and interfering speech are intelligible, leading ASR to misinterpret interfering speech. This is analogous to human distraction by irrelevant information, impairing their comprehension of target information. With fixed masking sound, target WER exceeds 0.65, indicating the need for selective sound masking and active target enhancement. AMasker achieves an average target WER of 0.0754, comparable to "TSE-only" (0.0703), representing over a 4-fold reduction compared to "No processing".

Comparison of interference suppression. Figs. 12 (b) & (c) compare interference suppression of baseline methods via SII/MSR of interfering speech. Since the target speech also masks the interfering speech, we compute the SII of "No processing" and "TSE-only" with the target speech as the masking sound; MSR, meaningless for the desired target speech, applies only to "Fixed masker" and AMasker. The SII of the "No processing" method (0.7286 on average) far exceeds the 0.2 threshold, indicating a lack of intelligibility suppression capability. The "Fixed masker" method exhibits unstable performance, with many samples at SII $>$ 0.2 and MSR either excessively high or excessively low. This is because a fixed masking sound cannot adapt to the temporal and spectral variations of interference, causing insufficient masking in certain regions. AMasker is the only method keeping SII below 0.2 (0.1270 on average, 75.8% lower than the "TSE-only" method), with an average MSR of 9.41 dB.

Subjective evaluation. To evaluate actual user experience, we reproduce three scenes in Fig. 11 (a) with 12 volunteers in place of the dummy participants. 10-minute trials are repeated for each method and scene until every volunteer has participated. Each then rates MOS as introduced in Sec. VII-B3. Per-scene results appear in Fig. 13 (b) $\sim$ (c). Averaged over the three scenes, "TSE-only" reaches an IS-MOS of only 1.67 and an Overall MOS of 2.31, indicating that most volunteers still find the interfering speech intelligible and distracting. AMasker raises IS-MOS to 3.44 and Overall MOS to 3.78. The "Fixed masker" matches our IS-MOS but scores lower on Overall MOS, as it suppresses the target speech as well.

Environmental generalization. To validate AMasker's generalization across different environments, we conducted experiments in 5 additional environments under the setting of Scene 2, as introduced in Sec. VII-B. The results in Fig. 14 show that the performance in Env. 1-5 is consistent with that in Sec. VII-C. Across all environments, the target speech WER stays below 0.08, while the SII and MSR remain below 0.2 and 13 dB, respectively.

D. Impact of spatial parameters

In this section, we evaluate AMasker under different spatial parameters shown in Fig. 15. All experiments are conducted in the environment shown in Fig. 11 (a), where we vary one parameter at a time while fixing the others to their Scene 2 defaults. Both groups follow the same setting, and all listeners are evaluated.

1) Target-listener distance: AMasker is robust to SNR variations at the listener caused by different target–listener distances, due to the active target enhancement strategy. To validate this, we vary this distance and test AMasker. As shown in Fig. 16, when the target-listener distance increases from 1 m to 2.5 m, all the metrics remain stable and within an acceptable range. Notably, the WER of the target speech consistently stays below 0.08.

2) Interference-listener distance: We further test AMasker across distances between the interfering speaker and the listener. With a selective sound masking mechanism, AMasker reduces the undesired intelligibility without hurting the target intelligibility. We varied this distance from 1 m to 2.5 m, and the results in Fig. 17 show that AMasker still maintains stable performance with SII lower than 0.2, MSR lower than 13 dB, and WER lower than 0.1.

3) Relative position between smartphone and listener: In practice, users may casually place smartphones around them. In this experiment, we vary both the orientation and distance between the smartphone and the listener, and test AMasker. The results in Fig. 18 validate AMasker's robustness under different relative positions, with WER, SII, and MSR remaining below the acceptable threshold.

E. Impact of practical factors

The environment and conversation setup in this section follow Sec. VII-D.

1) Non-intelligible noise: We play different types of non-intelligible noise through a speaker, varying the gap between the noise energy E_N and the interfering speech energy E_I at the listener. As Fig. 19 shows, AMasker is unaffected while E_N does not exceed E_I, since the target SNR at the target speaker's phone stays high and the separation model is trained to extract only intelligible components. Once E_N exceeds E_I by 10 dB, the reduced input SNR raises the average target WER to 0.1587. Meanwhile, SII drops, as the noise itself masks the interfering speech.

2) Mobility of users and devices: We test three mobility levels: *i) Static*, where participants and devices remain still; *ii) Moderate*, where the dummy participant is shaken to mimic head movements (e.g., nodding) and the device is displaced

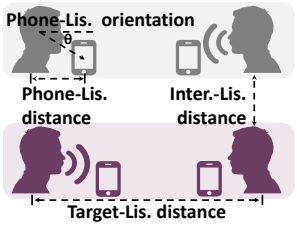


Fig. 15: Spatial parameters.

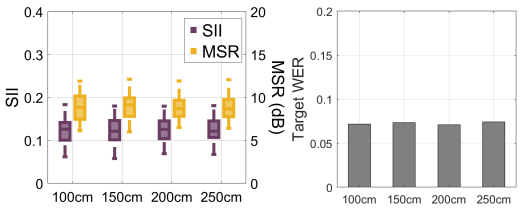


Fig. 16: Impact of target-listener distance.

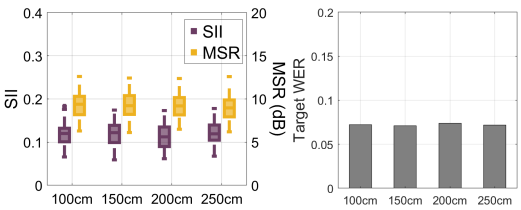


Fig. 17: Impact of interference-listener distance.

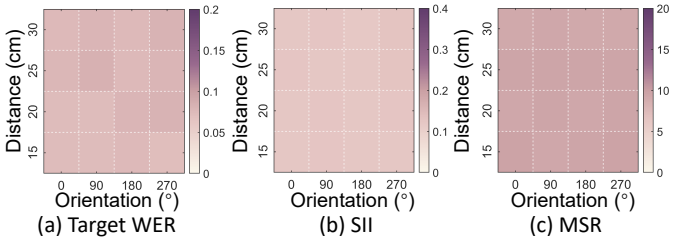


Fig. 18: Impact of phone-listener relative position.

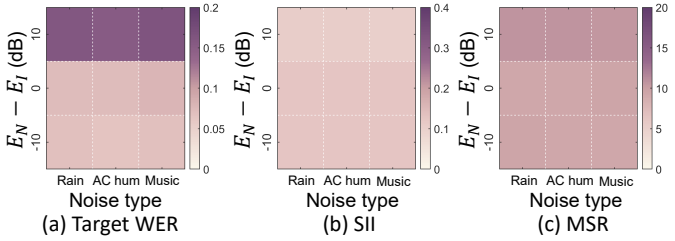


Fig. 19: Impact of non-intelligible noise.

by 50 cm every 30 seconds; *iii) Intensive*, where both the participant and the device move continuously within a 50 cm radius circle. The results in Fig. 20 (a)-(b) show that occasional device movement does not impair AMasker's performance, as the TDoA label can be re-established within a few hundred milliseconds. However, under intensive mobility, the target WER is above 0.13, as highly volatile TDoA measurements force speaker identification to rely entirely on the lightweight DNN model, reducing accuracy below 90% (Fig. 20(c)). This causes portions of the target speech to be misclassified as interfering speech and masked.

F. Time and power consumption

We profile the overhead of AMasker's key modules on a commodity smartphone, the Honor Magic 6 with a Snapdragon 8 Gen 3 SoC. Table II reports the average latency and memory usage over 100 runs of each module.

Online modules. The masking pipeline, including speech separation and masking sound generation, and the identification pipeline, including speaker identification and TDoA calculation, run in parallel. Each is dominated by the AI model it hosts: the separation and identification models account for 89.9% and 71.7% of their pipeline latency. Neither exceeds 76 MB of memory, modest for phones equipped with at least 6 GB of RAM. The masking pipeline takes 9.52 ms on average to process each 12-ms frame, a real-time factor (RTF) of 0.79, where RTF is the ratio of processing time to the interval between successive inputs. This latency is only 47.6% of the 20-ms budget that the 32-ms tolerance (Fig. 7) leaves after

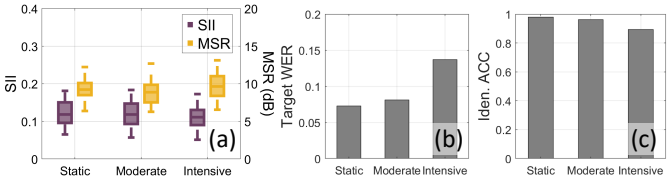


Fig. 20: Impact of mobility.

TABLE II: Overhead of AMasker's key modules.

Module	Type	Input (ms)	Latency (ms)	Memory usage (MB)
Iden. model	online	500	7.06 ± 0.50	74.21
Iden. pipeline	online	500	9.85 ± 0.53	75.09
Sep. model	online	12	8.56 ± 0.63	56.53
Masking pipeline	online	12	9.52 ± 0.91	59.08
Emb. model	offline	5000	3465.71 ± 158.95	963.11

the 12-ms window. The identification pipeline takes 9.85 ms to process each 25-ms hop (Sec. VI-A), corresponding to an RTF of 0.39. Since both RTFs stay well below 1, each pipeline consumes audio faster than it arrives, so no processing backlog accumulates during continuous operation.

Offline module. The user embedding acquisition module, which is not subject to real-time constraints, takes under 3.5 s on average and less than 1 GB of memory. The embedding module only runs once per user, since the extracted voiceprint can be reused across conversations.

Power consumption. AMasker's whole-device power consumption on the smartphone, measured by Android's BatteryManager APIs, is 3.6 W, allowing over 5 h of continuous operation on a 5000 mAh battery.

VIII. CONCLUSION

We present AMasker to enable a selective sound-masking workflow that enhances target intelligibility and precisely suppresses interfering intelligibility. It couples spectral geometric reshaping, which eliminates intelligibility leakage, with a collaborative speech separation scheme distributed across commodity devices. We implement a prototype and our objective and subjective results confirm that AMasker successfully creates a focused acoustic environment.

IX. USE OF AI DISCLOSURE

Generative AI image tools were used in two figures: OpenAI GPT-Image rendered the two-panel conceptual scene in Fig. 1 (a), and Gemini 2.5 Flash Image rendered the top-view vehicle sketch of Env. 5 in Fig. 11 (b). All authors inspected the generated artwork and confirmed that it faithfully depicts the setups described in the text.

REFERENCES

[1] A. Haapakangas, V. Hongisto, M. Eerola, and T. Kuusisto, "Distraction distance and perceived disturbance by noise—an analysis of 21 open-plan offices," *The Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 127–136, 2017.

[2] A. Haapakangas, V. Hongisto, J. Hyönä, J. Kokko, and J. Keränen, "Effects of unattended speech on performance and subjective distraction: The role of acoustic design in open-plan offices," *Applied Acoustics*, vol. 86, pp. 1–16, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:39508450

[3] W. Ellermeier and J. Hellbrück, "Is level irrelevant in "irrelevant speech"? effects of loudness, signal-to-noise ratio, and binaural unmasking," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 24, no. 5, pp. 1406–1414, 1998.

[4] A. Gupta, "Sound masking system market size, industry report, 2024–2032," https://www.marketresearchfuture.com/reports/sound-masking-system-market-8145, 2019, report ID: MRFR/ICT/6673-CR. Market Research Future. Accessed: 2026-03-04.

[5] L. Brocolini, E. Parizet, and P. Chevret, "Effect of masking noise on cognitive performance and annoyance in open plan offices," *Applied Acoustics*, vol. 114, pp. 44–55, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X16302067

[6] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Brill, 2012.

[7] "Sound masking systems," https://www.pureresonanceaudio.com/collections/sound-masking-systems, Pure Resonance Audio, accessed: 2025-01-28.

[8] "Elite acoustics product line," https://www.elite-acoustics.com, Elite Acoustics, accessed: 2025-01-28.

[9] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *The Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[10] V. Dellwo, A. Leemann, and M.-J. Kolly, "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1513–1528, 2015.

[11] T. Renz, P. Leistner, and A. Liebl, "Auditory distraction by speech: Sound masking with speech-shaped stationary noise outperforms- 5 db per octave shaped noise," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. EL212–EL217, 2018.

[12] B. Veluri, M. Itani, T. Chen, T. Yoshioka, and S. Gollakota, "Look once to hear: Target speech hearing with noisy examples," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–16.

[13] T. Chen, M. Itani, S. E. Eskimez, T. Yoshioka, and S. Gollakota, "Hearable devices with sound bubbles," *Nature Electronics*, pp. 1–12, 2024.

[14] Lencore, "Case study: Knight Ridder," https://www.lencore.com/case-studies/case-study-knight-ridder/, n.d., vendor case study. Accessed: 2026-03-14.

[15] J. W. Newbold, J. Luton, A. L. Cox, and S. J. Gould, "Using nature-based soundscapes to support task performance and mood," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, pp. 2802–2809.

[16] Haworth, "Acceptance & Efficacy of Biophilic Soundscaping in An Open-Plan Office," Haworth, Research Report, 2021, accessed on July 7, 2024.

[17] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer Science & Business Media, 2013, vol. 22.

[18] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.

[19] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1," p. 27403, 1993.

[21] T. H. Crystal and A. S. House, "Segmental durations in connected-speech signals: Current results," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1553–1573, 1988.

[22] J. A. Veitch, J. S. Bradley, L. M. Legault, S. Norcross, and J. M. Svec, "Masking speech in open-plan offices with simulated ventilation noise: noise level and spectral composition effects on acoustic satisfaction," *Institute for Research in Construction, Internal Report IRC-IR-846*, 2002.

[23] *Methods for Calculation of the Speech Intelligibility Index*, American National Standards Institute Std. ANSI S3.5-1997, 1997.

[24] "Criteria for a Recommended Standard: Occupational Noise Exposure, Revised Criteria 1998," National Institute for Occupational Safety and Health (NIOSH), Tech. Rep., 1998.

[25] K. S. Pearsons, R. L. Bennett, and S. A. Fidell, "Speech levels in various noise environments," Office of Health and Ecological Effects, Office of Research and Development, US Environmental Protection Agency, Tech. Rep., 1977.

[26] M. R. Schroeder, B. S. Atal, and J. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *The Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

[27] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[28] M. Kazama, S. Gotoh, M. Tohyama, and T. Houtgast, "On the significance of phase in the short term fourier spectrum for speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1432–1439, 2010.

[29] H.-T. Chiang, K.-H. Hung, S.-W. Fu, H.-C. Kuo, M.-H. Tsai, and Y. Tsao, "Study on the correlation between objective evaluations and subjective speech quality and intelligibility," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[31] *Acoustics — Audiometric Test Methods — Part 3: Speech Audiometry*, International Organization for Standardization Std. ISO 8253-3:1996, 1996.

[32] S. Yue, P. Pilon, and G. Cavadias, "Power of the mann–kendall and spearman's rho tests for detecting monotonic trends in hydrological series," *Journal of hydrology*, vol. 259, no. 1-4, pp. 254–271, 2002.

[33] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[34] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, 2002.

[35] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in psychology*, vol. 6, p. 136034, 2015.

[36] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn taking for conversation," in *Studies in the organization of conversational interaction*. Elsevier, 1978, pp. 7–55.

[37] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[38] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[39] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[40] Z. Zhang, C. Chen, H.-H. Chen, X. Liu, Y. Hu, and E. S. Chng, "Noise-Aware Speech Separation with Contrastive Learning," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 1381–1385. [Online]. Available: https://ieeexplore.ieee.org/document/10448214/

[41] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[42] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.