# Hear What You Want: Headphone-free Sound Bubbles via Selective Sound Masking

Huangwei Wu*, Yilin Yao*, Meng Jin†, Weiguo Wang†, Qianwu Chen, Tingchao Fan, Xinbing Wang, Chenghu Zhou

Shanghai Jiao Tong University, ByteDance, Chinese Academy of Sciences

## ABSTRACT

Conversations in shared spaces often cause inter-group interference, which distracts the participants. While noise-canceling headphones mitigate this, they remain costly and uncomfortable for ubiquitous adoption. We present *AMasker*, a novel system that can mitigate the interfering speech without specialized hardware. AMasker operates on the insight that the distraction caused by interfering speech stems from its intelligibility, which can be suppressed by adding a delicately designed masking sound, thereby avoiding complex cancellation. Combining advanced sound masking and speech separation technologies, AMasker can selectively mask interfering speech while enhancing target speech. We implement a prototype on smartphones 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 on general audio devices.

## 1 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 [1–3].

Existing works [4, 5] offer a potential solution by extracting the target speech from the mixture and replaying it to the user. Meanwhile, the interfering speech is suppressed using the Active Noise Cancellation (ANC) technique, which generates an anti-noise signal that is the exact "opposite" of the interfering signal, thereby canceling it out. Yet, one problem underlying ANC is that propagation of the anti-noise signal in the environment may introduce unpredictable signal distortion and delay. Even cm-level propagation delays can cause misalignment between the interfering signal and the anti-noise (see Fig. 2 (a)), which fails the noise cancellation process [6, 7]. So, existing ANC techniques typically require the user to wear a specialized headphone, which plays an anti-noise signal in the user's ear to minimize uncertainty in the signal's propagation delay and distortion.

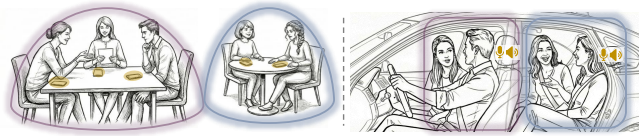


**Figure 1: Bubbles of intelligibility created by AMasker with general audio devices.**

However, the reliance on noise-canceling headphones severely limits the ubiquitous adoption of ANC-based methods. First, the cost barrier limits the widespread adoption of noise-canceling headphones – the market penetration of ANC headphones is less than 3% in 2024 and is expected to be under 12% even by 2034 [8, 9]. Furthermore, prolonged headphone use can be uncomfortable and inconvenient, potentially leading to allergic contact dermatitis (ACD) from skin contact with sensitizers in headphone materials [10, 11].

We explore the following question: *Could we mitigate the interference of undesired speech without relying on a headphone?* For example, can we use users' personal devices (e.g., smartphones) or existing infrastructure (e.g., in-vehicle audio systems) to mitigate inter-group interference in shared spaces? We achieve this based on an observation: the distraction from background noise is more determined by its *intelligibility*, rather than its loudness [12, 13]. Existing studies [14] show that compared with random noise (e.g., wind, rain), intelligible speech is more likely to distract individuals' attention. Therefore, improving employees' performance does not require fully canceling the distracting speech; reducing its intelligibility is also a promising solution.

Sound masking is a viable method for intelligibility suppression [15–17]. It is derived from a well-known psychoacoustic phenomenon - **masking effect**. Specifically, due to the characteristics of the human auditory system, the perception of speech will be reduced by the presence of another louder masking sound that is played in close temporal proximity [18]. Generally, the more the spectrum of the masking sound overlaps the masked speech, the more effective the

* Co-first authors. †Co-corresponding authors.

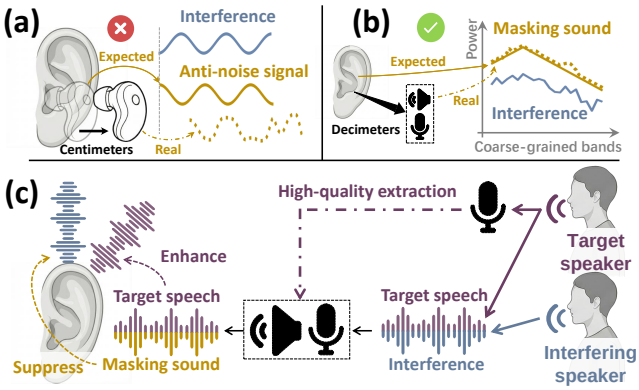



**Figure 2: Contrasting (a) precise waveform cancellation (ANC) with (b) coarse-grained spectral masking, motivating (c) AMasker's selective masking design.**

masking becomes [19]. Numerous studies reveal that non-intelligible masking sounds (such as pink/white noise) can significantly mitigate the interference of intelligible speech and improve work efficiency even when their sound pressure levels are higher than those of intelligible speech [20]. This phenomenon happens in everyday life. For example, a whispering speech that is noticeably distracting in a quiet library will become imperceptible during heavy rainfall. However, although commercial sound masking systems are gaining increasing market share [21–23], they commonly employ fixed masking sounds to broadly reduce the intelligibility of *all speech* in the environment, including the target speech that the user is trying to focus on. We need a **selective sound masking** method to mask only the undesired speech with minimal negative impact on the target speech.

In this paper, we propose AMasker, a novel system that introduces *Active Intelligibility Control* (AIC), a simple yet effective technique that generates highly customized masking sounds to achieve the desired selective masking. Our *key insight* is that, due to the difference in speech contents and the speakers' voice patterns, the target speech and the interfering speech exhibit substantial asynchrony in both temporal waveforms and spectral characteristics. Given that sound masking depends critically on spectral alignment between the masking signal and the masked speech, AMasker achieves selective sound masking by generating the masking sound that aligns specifically with the interfering speech. Compared with ANC, AMasker's selective sound masking imposes less stringent hardware requirements, as its effectiveness relies solely on coarse-grained spectral alignment between the masking sound and the interfering speech (see Sec. 5 for details), rather than on precise waveform-level matching. As illustrated in Fig. 2 (b), the masking performance remains robust even when the masking sound suffers a dm-level propagation delay and distortion (see Sec. 3.3).

Fig. 2 (c) illustrates the basic idea to design AMasker, which leverages collaboration among users' audio devices (e.g., their smartphones) to create *bubbles of intelligibility* as illustrated in Fig. 1 — a mutually masked environment — for conversation groups. The bubbles simultaneously eliminate distractions and improve the intelligibility of target speech. To achieve this, AMasker first extracts the target speech and the interference speech from the mixture. Then it generates an intelligibility control signal, which is a superposition of the separated target speech for target enhancement and a masking sound delicately designed based on the interference speech for interference suppression. Note that the target speech is extracted at the speaker's device and relayed wirelessly to the listeners' devices since the speaker's device captures the speech with a higher SNR. While the interfering speech is extracted at the listener's device to obtain a signal that is closer to what the listener actually hears. AMasker is a software-only ubiquitous system, which can be deployed on any general device with a speaker and microphone. Three challenges arise in implementing AMasker.

**Challenge 1:** *Suppressing intelligibility of interfering speech while keeping low loudness.*

An effective masking sound should satisfy two requirements: while suppressing the intelligibility of interfering speech, its own intensity should remain below a certain threshold (85 dBA) to minimize the harm to the user's auditory system. One naive solution is to align the masking sound's spectrum closely with the interference spectrum and set its volume slightly higher than that of the interference. However, this makes the spectrum of the masking sound too similar to that of the interfering speech, leading to "intelligibility leakage", that is, the masking sound *"emulates"* the intelligibility of interfering speech, making the masking sound itself intelligible.

We address the above issue with a spectral geometric reshaping method. This method generates masking sound that preserves the coarse-grained spectral trends of the interference while deliberately smearing the intelligibility-critical fine-grained details of the sound wave.

**Challenge 2:** *Target speech separation in multi-speaker scenarios.*

Since perfect time–frequency misalignment between the target and interfering speech is unattainable, the masking sound will inevitably suppress the target to some extent. To counteract this, AMasker first extracts the target speech from the mixture and replays it to the listener. However, most existing target speech separation methods are limited to separating speech of a single, fixed speaker, making them unsuitable for dynamic multi-speaker scenarios. These approaches typically require prior knowledge of the target speaker's voiceprint—represented as a speaker embedding extracted from a quiet, second-scale voice recording. Yet,

in multi-speaker environments, the speaker switches frequently among participants. It is difficult to obtain all the speakers' embeddings in advance.

To address this problem, we propose a continual voiceprint extraction method that eliminates dedicated enrollment by automatically extracting embeddings from clean speech acquired during users' daily interactions with personal devices (e.g., smartphones). During conversations, users' voiceprints are shared via the local network to build a voiceprint table on each user's smartphone. We then propose a collaboration-based speaker identification technology that combines an identification model with spatial cues (i.e., TDoA) to determine the active speaker robustly in real time.

**Challenge 3:** *Interfering speech separation without the information of voiceprint.*

To generate masking sounds, we need to further isolate the interfering speech from the mixed signal. Intuitively, one could obtain the interfering signal by subtracting the target speech from the mixed signal. However, the extracted target speech, although it retains intelligibility, its waveform will inevitably undergo distortion. So, subtracting the target speech from the mixed signal corrupts the estimated interfering speech. Directly extracting the interfering signal is also challenging, because the interfering signal typically consists of speech from multiple speakers. It is nontrivial for current speech separation methods to restore overlapping speech from multiple speakers, without their voiceprints.

We address the above challenge with a dual-decoder speech separation architecture which treats the interfering speeches as a whole and extracts the target and non-target speech simultaneously from the mixture. By leveraging the target speaker's voiceprint to delineate the feature boundary between the target and non-target speeches, the model is able to reconstruct the non-target portion with improved fidelity.

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× compared with baselines lacking 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, achieving the highest scores among headphone-free baselines. It also meets real-time requirements, with a processing delay below 50% of the maximum latency tolerance. We summarize our contributions as:

- We propose the first system that leverages the intelligibility suppression to mitigate distracting background speech, without relying on headphones.

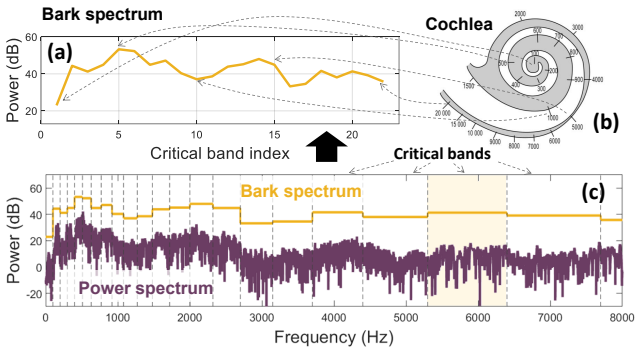



**Figure 3: Power spectrum and Bark spectrum.**

- We design a selective sound masking method that selectively reduces the intelligibility of interference to mitigate distracting speech, while enhancing the target speech.
- We present a working prototype of AMasker and validate its effectiveness in extensive real-world scenarios.

## 2 RELATED WORK

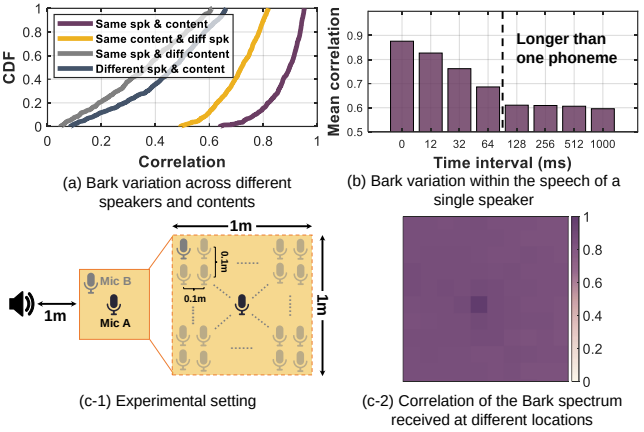
**Headphone-based selective hearing.** Recent efforts in headphone-based selective hearing have shown promising results, enabling users to focus on a target speaker amid interfering speech. For instance, [4] proposed a "look once to hear" approach, using visual input to guide target speech hearing in noisy environments. Their system relies on ANC to suppress interference, necessitating specialized ANC-capable headphones. Similarly, [5] introduced the concept of "sound bubbles" for hearable devices, leveraging ANC to create localized listening zones around the user. While effective, these methods share a common requirement: head-worn hardware with integrated ANC. In contrast, AMasker achieves selective hearing without requiring any worn device, offering a more flexible and accessible solution.

**Sound masking systems.** Sound masking has attracted considerable attention and recognition in academia [12, 13, 19, 24] and has been successfully deployed in numerous real-world office environments (e.g., renowned companies such as Knight Ridder [25] and Cisco [26]), receiving highly positive evaluations for its noise-suppression performance. However, most existing approaches rely on stationary masking signals—ranging from conventional broadband noise to natural sounds [27–29]—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.

## 3 BACKGROUND AND INTUITION

### 3.1 Sound masking effect

The masking effect is a psychoacoustic principle where the perception of one sound (the "maskee") is reduced by another louder sound (the "masker"). This originates from




(a) Bark variation across different speakers and contents

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

(c-1) Experimental setting

(c-2) Correlation of the Bark spectrum received at different locations

**Figure 4: Analysis of Bark spectrum variations across different conditions.**

the cochlea's frequency-resolving mechanism. When two sounds are close in frequency, they excite overlapping regions on the basilar membrane, suppressing the neural response to the weaker "maskee". Moreover, the auditory system acts as a set of bandpass filters whose bandwidth—the *critical band*—reflects the cochlea's frequency resolution [30]. Sounds within the same critical band are processed as a single entity, making them hard to resolve individually. Harnessing this principle, sound masking sends a controlled masking sound [1] to reduce the intelligibility of interfering speech.

## 3.2 Bark spectrum

AMasker generates the most effective masking sound by analyzing the Bark spectrum of the interfering speech. A Bark spectrum is a spectral representation of an acoustic signal. Compared with other spectral representations (e.g., Mel spectrum), the Bark spectrum utilizes a frequency scale that more closely matches human auditory perception [31, 32]. Specifically, the Bark spectrum divides the audible frequency range (roughly 20 Hz to 20 kHz) into 24 critical bands, within which sounds are processed together by the auditory system. The power of each critical band $i$ can be calculated as:

$$B(i) = \sum_{k=bl(i)}^{bh(i)} P[k] \tag{1}$$

where $bl(i)$ and $bh(i)$ are the lower and upper frequency boundaries of $i$-th critical band, and $P[k]$ is the power of $k$-th frequency bin. Fig. 3 shows an example Bark spectrum and corresponding power spectrum. The critical bands are unevenly spaced on the linear frequency scale, with bandwidth increasing with frequency because human auditory resolution is coarser at higher frequencies.

## 3.3 Feasibility study

The Bark spectrum reflects how the human auditory system perceives sound across critical bands. In this section, we examine speech perception from the Bark-spectrum perspective across speakers/content, over time, and across spatial positions. We demonstrate key properties of the speech Bark spectrum that validate the feasibility of selective sound masking and guide AMasker's deployment.

**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 [33] (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. 4 (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. 4 (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 [34]).

**Property 3:** *Bark-spectrum characteristics of speech exhibit limited variation under channel effects*

Because a user's smartphone and ears are spatially separated, interfering speech reaches them via different acoustic channels. Consequently, the interfering speech perceived by the listener may differ spectrally from what is captured at the device, potentially limiting the effectiveness of masking sounds generated from the device's microphone.

To investigate whether this positional discrepancy induces noticeable differences at Bark scale, as shown in Fig. 4 (c-1), we conducted experiments in a 1m×1m central area of the room shown in Fig. 12 (a). A speaker is placed 1 m from the area, and two microphones (Mic A and Mic B) record its output. Mic A is fixed at the center while Mic B was moved across different positions. A 1-minute speech clip is played and recorded by both microphones, and then the Bark spectrum of the received signals is computed. Fig. 4 (c-2) visualizes the correlation between the two microphones' Bark spectra for various Mic B positions. The Bark correlation remains above 0.8 when the inter-microphone distance is less than 0.5 m.

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

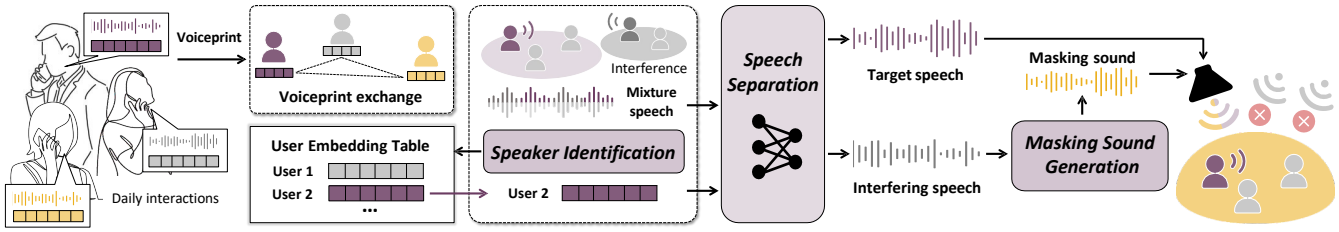



**Figure 5: Overview of AMasker.**

This indicates that, despite multipath effects in indoor environments, the human auditory system's limited frequency resolution—particularly at high frequencies—results in minimal perceptual impact from such channel-induced spectral variations. Sec. 7 further confirms that channel effects do not significantly degrade masking performance.

**Insights.** Our Bark spectrum analysis yields three key conclusions for system design: i) human speech contains sufficient spectral diversity for auditory distinction, making it feasible to generate masking sound that selectively masks a specific speech; ii) masking sounds must adapt in real time to the time-varying spectral characteristics of interfering speech; and iii) it's feasible to achieve selective sound masking at positions away from the listener's ear in a wearable-free manner.

## 4 SYSTEM OVERVIEW

AMasker realizes active intelligibility control by selectively reducing the intelligibility of interfering speech and enhancing that of the target speech. Fig. 5 summarizes the system architecture, which consists of three core components and operates in three stages. In the first stage, user embeddings are collected during daily smartphone interactions (e.g., phone calls), where the system records voiceprint features. Before a conversation begins, participants exchange embeddings to construct a shared embedding table. During the conversation, devices collaboratively identify the current speaker, retrieve the corresponding voiceprint from the table, and the speaker separation module extracts the target and interfering speech. The system then generates a masking sound to suppress the interference and replays the target speech to enhance it.

Next, we describe how to generate effective masking sounds from extracted speeches (see Sec. 5), followed by the speech separation method supporting this process (see Sec. 6).

## 5 MASKING SOUND GENERATION

The objective of sound masking is to suppress the intelligibility of interfering speech, as evaluated by the Speech Intelligibility Index (SII). SII is an index ranging from 0 to 1 that quantifies the intelligibility of an interfering sound $S_I$ in the presence of a masking sound $M$. A smaller SII indicates lower intelligibility. SII reflects a weighted sum of the energy gap between $M$ and $S_I$ in each critical band. It can be calculated as defined in the ANSI S3.5-1997 standard [35] as:

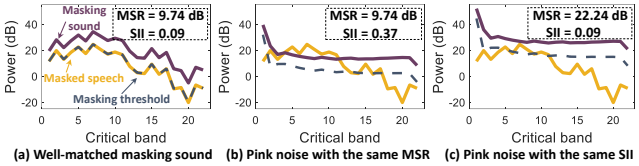



**Figure 6: Impact of masking threshold's alignment.**

$$SII = \sum_i I_i \cdot \max\left(0, \min\left(1, \frac{B_{Ii} - D_i\left(B_M\right) + 15}{30}\right)\right) \quad (2)$$

where $I_i$ is an empirically defined importance coefficient of the $i$-th critical band, which reflects human auditory sensitivity to that band [35]. We have $\sum I_i = 1$. $B_I$ is the Bark spectrum of the interfering speech. $D(B_M)$ is an empirically defined mapping that transforms the masking sound's Bark spectrum ($B_M$) into a perceptually refined energy level that describes its masking capability. A larger $B_M$ yields a higher $D(B_M)$, indicating better masking performance (and thus a smaller SII). Eq. (2) tells that when $D(B_M)$ exceeds the interfering speech energy by 15 dB across all critical bands, the interfering speech becomes totally unintelligible (with $SII = 0$). Studies also show that interfering speech with SII higher than 0.2 can cause distraction [36].

As indicated in Eq. 2, continuously increasing masking sound energy can ultimately reduce SII below a satisfactory threshold. However, excessive masking sound (e.g., above 85 dBA) can damage the human auditory system [37]. Moreover, although the target and interfering speech have quite different spectral characteristics, they cannot be completely isolated in either the time or frequency domain. Without loudness control, the masking sound will inevitably reduce the target speech intelligibility. So, we should also minimize the masker-to-signal ratio (MSR) of the masking sound:

$$MSR\left(M, S_I\right) = 10 \log 10 \left(\frac{Sum(P\left(M\right))}{Sum(P\left(S_I\right))}\right) \quad (3)$$

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 below 0.2) with minimal loudness (i.e., MSR below 20 dB, given that human speech is typically 55-65 dBA [38]).

**The ideal masking sound.** An interfering speech is well masked when its Bark spectrum is under the **masking**

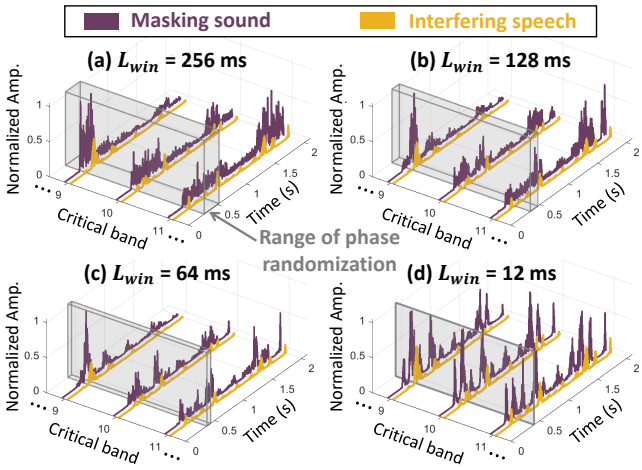



**Figure 7: Comparison of spectro-temporal structures across different processing window lengths.**

**threshold** of a masking sound [39, 40], which is a Bark-scaled energy level across critical bands below which the interfering speech becomes unperceived. Existing studies have developed the forward calculation from a masking sound's power spectrum to its masking threshold [41, 42]. To maximize masking efficiency, the ideal strategy is to align the masking sound's spectrum so that its masking threshold precisely covers the interfering speech's Bark spectrum. Fig. 6 illustrates this concept: a well-aligned masking sound achieves effective intelligibility suppression (SII=0.09) while maintaining a low loudness (MSR=9.74 dB). Thus, to obtain the ideal masking sound $M^*$, we can treat the interfering speech's Bark spectrum as the masking threshold and solve the inverse problem. The details are shown in Appendix A.

## 5.1 Intelligibility leakage problem

One critical flaw of the above method is that, when aligning the Bark spectrum of the masking sound closely with that of the interference, the spectral envelopes of the two sounds are too similar, leading to the "intelligibility leakage" issue: the masking sound *emulates* the content of the interfering speech, making the masking sound itself intelligible.

An effective method to address the intelligibility leakage problem is adding randomness to the phase of $M^*$. Specifically, *speech content is encoded by spectro-temporal cues across narrow frequency bands* [43–46], arising from two joint 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 temporal envelope fluctuations and destroying intelligibility. Fig. 7 (a) shows the temporal envelopes of $M^*$ in specific critical bands (920-1480 Hz), with their phase scrambled within 256-ms windows. We can see that temporal envelopes of $M^*$ are severely blurred compared to those of the interfering speech, preventing intelligibility leakage.

**Table 1: WER of different masking sounds.**

| 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 |

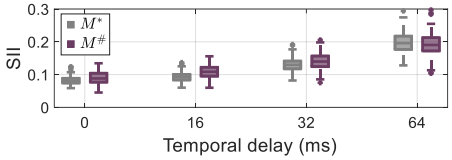



**Figure 8: Impact of temporal delay on masking.**

However, the above phase scrambling method is effective only when the processing window length ($L_{win}$) is sufficiently long. This is because when we use a short $L_{win}$, the limited scrambling range is insufficient to smear the rapid rhythmic changes (e.g, the sharp peaks and valleys in temporal envelopes), preserving the intelligibility-critical cues. Figs. 7 (b)-(d) show temporal envelopes of the ideal 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 [47, 48], we further use the SOTA model Whisper-large-v3 [49] to evaluate the intelligibility of $M^*$ generated with different $L_{win}$. The results in Table 1 indicate that $L_{win}$ should be at least 128 ms to keep $M^*$ unintelligible (i.e., WER > 0.5 [50]).

However, under real-time constraints, a short processing window is necessary as longer windows introduce higher latency, causing temporal asynchrony between the masking sound and interfering speech and thus unacceptable performance degradation. To quantify this impact, we generated masking sounds for 40 speakers from the LibriSpeech dataset, applied varying temporal delays, and computed the SII of the masked speech (using the method introduced in Sec. 5). As shown in Fig. 8, while the tolerance is considerably higher than the $\mu$s-level alignment required for ANC [51], the masking sound delay should not exceed 32 ms to maintain SII below the 0.2 threshold. Based on the results in Sec. 7.6, we set the processing window length $L_{win}$ to 12 ms, yielding an average processing latency of 9.52 ms.

## 5.2 Spectral geometric reshaping

To solve the intelligibility leakage problem, beyond time-domain approaches (i.e., phase randomization) incompatible with real-time processing, we can also operate in the frequency domain by fine-tuning the Bark spectrum of the

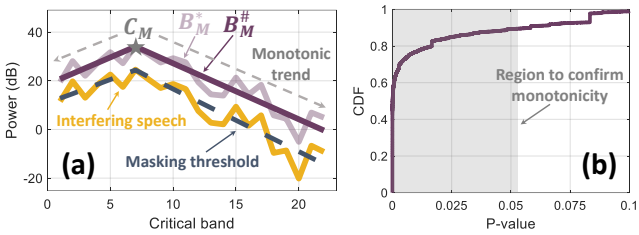



**Figure 9: (a) Illustration of the proposed adaptive Bark reshaping method; (b) Results of monotonicity test.**

ideal masking sound $B^*_M$, making its masking threshold fully cover the spectrum of interfering speech, while reducing its spectral similarity with the interference. However, such a multi-objective optimization problem is traditionally difficult because i) it depends on iteration-based methods (i.e., gradient descent), incurring catastrophic computational overhead; ii) it requires a guiding metric, while for real-time processing, no metric can evaluate ms-level audio's intelligibility.

**Insight.** The inherent spectral structure of speech in the Bark domain reveals a simple geometric transformation that bypasses complex optimization while achieving all objectives simultaneously. Specifically, we observe that although the speech Bark spectrum varies with speech content, *its shape exhibits a special "pyramid-like" pattern.* As shown in Fig. 9 (a), the speech Bark spectrum is split into two parts at the critical band with maximum energy, denoted as $C_M$. The Bark bins on both sides exhibit a monotonic trend. To further test this observation's reliability, we leverage Spearman's rho test [52] to examine the monotonicity of Bark bins split by $C_M$ across the LibriSpeech dataset. Given a sequence of Bark bins, this test returns a $p$-value representing the statistical significance of monotonicity, with $p < 0.05$ confirming it. Fig. 9 (b) shows that the $p$-values of nearly 90% testing samples are lower than 0.05. This reveals that with only two linear fittings, we can smear the intelligibility-critical fine-grained details in the spectral envelope while still preserving promising masking efficiency by maintaining similar coarse-grained spectral trends.

**Adaptive Bark reshaping.** We reshape $B^*_M$ based on the "pyramid-like" pattern of the interfering speech. Specifically, after specifying $C_M$ as a boundary point, we fit lines to the Bark bins on both sides with the least squares method (LSM). The combination of these two lines, which intersects at $C_M$, constitutes the new Bark spectrum $B^\#_M$, as shown in Fig. 9 (a). We further scale the Bark spectrum $B^\#_M$ to make its loudness equal to that of the original one (i.e., $B^*_M$). Then the new Bark spectrum $B^\#_M$ could further be leveraged to reconstruct the new power spectrum $P^\#(M)$ and the masking sound $M^\#$.

**Result.** With the same setting as others in Fig. 8, the impact of temporal asynchrony on $M^\#$ is also demonstrated in the figure. We can see that SII measurements of $M^\#$ are comparable to those of $M^*$. Furthermore, the WER results in

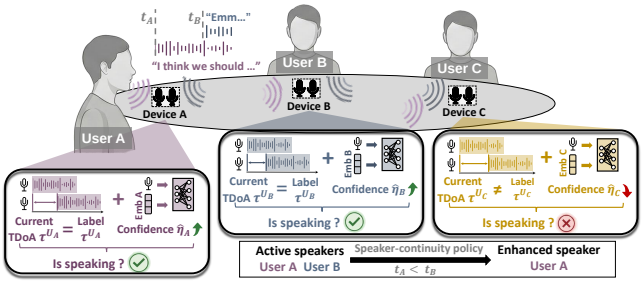



**Figure 10: The collaborative identification scheme.**

Table 1 show that $M^\#$ is more unintelligible than $M^*$ generated with a 256-ms phase randomization window.

## 6 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 [53].

**Embedding acquisition**. AMasker employs a lightweight embedding model suitable for resource-constrained devices [4], 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, we leverage the distinct spatial configuration of users' mobile devices. Concretely, 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 [54]. The separated target speech is then relayed to all listeners via the local network. Interfering speech separation is handled on each listener's device, as the received interfering signal better reflects 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.

### 6.1 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. 10, 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 selects the appropriate speaker embedding for subsequent processing.

**Speaker identification with hybrid cues.** Speaker identification must operate on sub-second speech windows to avoid speaker mixing, 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:

$$S_i(t) = \alpha(t)\hat{\eta}_i(t) + (1 - \alpha(t))\phi\left(\left|\tau_i(t) - \tau_L^{U_i}\right|\right) \quad (4)$$

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. $\phi$ 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, $\gamma_{U_i}$ 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,

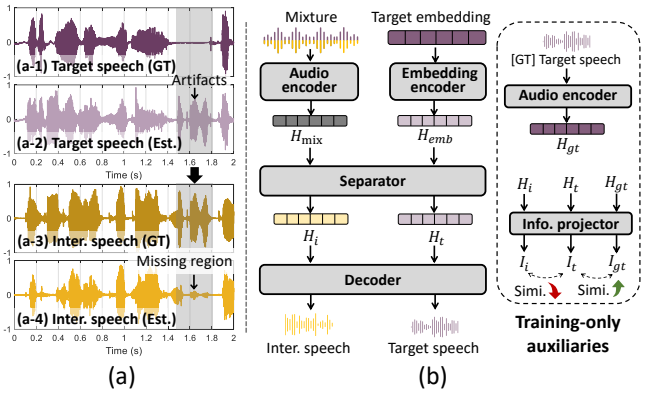


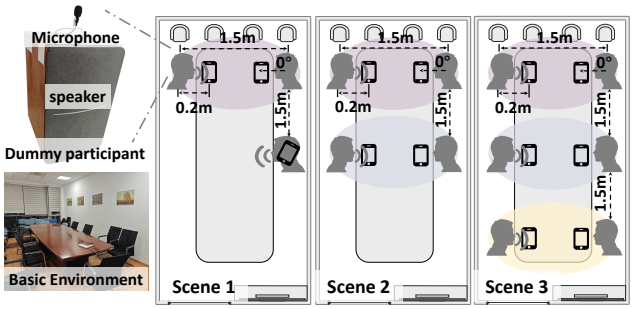

**Figure 11: (a) Example of error propagation (SNR=0 dB); (b) The dual-decoder architecture.**

where most overlaps are transient (lasting only a few hundred milliseconds [55]) and do not immediately alter the current dominant speaker [56].

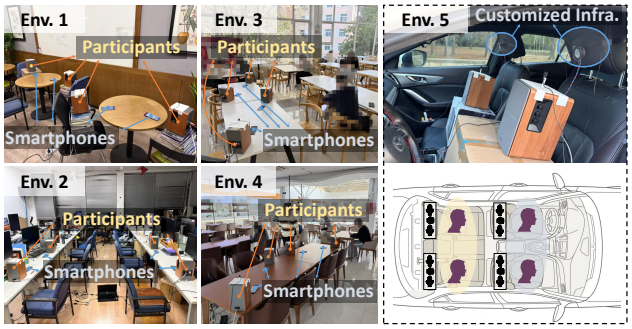
## 6.2 Accurate interfering speech separation

To obtain the interfering speech for masking sound generation, a naive method is to first separate the target speech using the target speaker's embedding, then subtract it from the mixture to obtain the interfering speech. However, this method relies heavily on accurate reconstruction of the target speech at the listener, which is difficult for separation models available on resource-limited devices under low SNR conditions [57]. As illustrated in Fig. 11 (a), when operating with low SNR, even a SoTA model [4] may incorrectly identify portions of interfering speech as target speech, resulting in the reconstructed target speech containing portions of the interfering signal (Fig. 11 (a-2)). This causes *error propagation*, where missing components in the reconstructed interference (Fig. 11 (a-4)) yield an ineffective masking sound.

**Dual-decoder architecture.** To solve this problem, we design a model architecture that explicitly separates both target and interfering speech with only the target embedding. This, on the one hand, avoids the error propagation that occurs in the subtraction-based method; on the other hand, it improves the separation quality for both the target and interfering speech. Specifically, this design explicitly enforces the model to disentangle interfering speech from the target speaker's representation space, thereby associating the target embedding only with the target speech features, preventing the accidental inclusion of interfering patterns. As shown in Fig. 11 (b), we adopt the widely accepted encoder-separator-decoder framework [58]. 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 $H_i$.




(a) Conversation setup




**(b) Diverse environments for test: workplaces (Env. 1 & 2), crowded public places (Env. 3 & 4), and an in-vehicle environment (Env. 5)**

**Figure 12: Experiment settings.**

**Training.** To further enhance the model's feature disentanglement ability, alongside the standard SNR loss [59], we adopt a contrastive learning loss $L_{cl}$ [60] to maximize the similarity between the latent features of the target speech and its ground truth, while minimizing the similarity between the target and interfering features. The final loss is:

$$L = \beta_1 \cdot SNR(s_t, \hat{s}_t) + \beta_2 \cdot SNR(b_I, \hat{b_I}) + \beta_3 \cdot L_{cl} \quad (5)$$

where $\hat{s}_t$ and $s_t$ are the estimated target speech and its ground truth. $\hat{b_I}$ and $b_I$ are the Bark spectrum of estimated interfering speech and its ground truth. Since masking only requires Bark signals, we constrain interference reconstruction to the Bark domain, simplifying internal feature mapping and improving quality. $\beta_1$, $\beta_2$, and $\beta_3$ are adjusted by an adaptive scaling method [61].

**Handling cross-device interference.** In real-world scenarios, speech undergoes both channel-induced distortion and interference from other masking sounds. To ensure that AMasker remains robust to these effects, we explicitly simulate these effects during training by convolving speech with diverse Room Impulse Responses (RIRs) and adding unintelligible noise (e.g., white noise and synthetic masking sounds). This augmentation steers the model to extract only intelligible components and prevents it from mistaking nearby devices' masking sounds as interference. This avoids the positive feedback loop in which devices continuously raise their masking-sound levels in response to each other.

# 7 EVALUATION

## 7.1 Implementation

In this paper, we primarily prototype AMasker using commercial smartphones from mainstream brands, including Honor, Google Pixel, Samsung, Xiaomi, Redmi, and iQOO. The signal is sampled at 48 kHz for accurate TDoA estimation and downsampled to 16 kHz for other tasks to reduce overhead. After the target speech is extracted, it will be amplified to maintain a 3 dB gain over the generated masking sound, then sent to the listener. We implement the collaboration network over Wi-Fi, using UDP for real-time transmission.

## 7.2 Experimental methodology

*7.2.1 Experiment settings.* We perform both objective and subjective evaluations. For objective evaluation, we replace humans with dummy counterparts consisting of commercial speakers and microphones, as shown in Fig. 12 (a), enabling reproducible, real-world synthesized conversations and precise metric computation. Each dummy participant is paired with a smartphone. As shown in Fig. 12 (a), conversations are conducted across three predefined scenes within a basic environment, with interference levels progressively increasing from Scene 1 to 3. Fig. 12 (b) shows five unseen environments used for test: typical workplaces (Env. 1 & 2), crowded public places (Env. 3 & 4), and an in-vehicle setting (Env. 5). For Env. 5, we customized an audio system with commercial microphones and speakers as the existing infrastructure. For subjective evaluation, approved by our IRB, 12 volunteers (7 males and 5 females) are invited to form groups and participate in conversations, as shown in Fig. 14 (a).

*7.2.2 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 [55], we insert random 100–400 ms overlaps at turn boundaries and mid-turn intrusions 300–500 ms—which cumulatively account for 5% of the total conversation duration.

*7.2.3 Metrics.* Comprehensive metrics are adopted as:

i) *Word Error Rate (WER)*: We calculate the ASR WER of the target speech in the same way as in Sec. 5.1. WER below 0.1 is considered acceptable for human understanding [62, 63].

ii) *Speech Intelligibility Index (SII)*: We calculate SII as introduced in Sec. 5. Note that SII should stay below 0.2.

iii) *Masker-to-Speech Ratio (MSR).* We calculate MSR as introduced in Sec. 5, and it should be lower than 20 dB.

For subjective evaluation, we collect mean opinion score (MOS) from volunteers to evaluate the ability of regulating listener's attention towards 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*

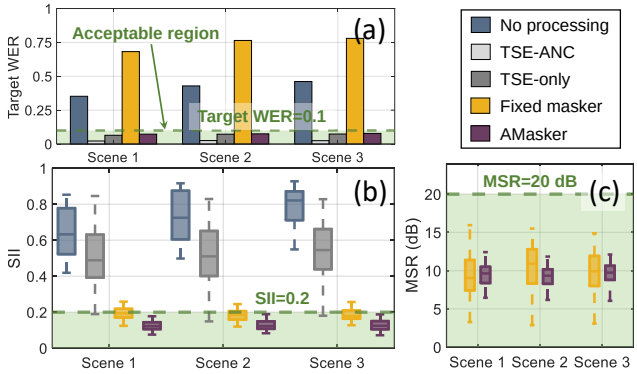



**Figure 13: Comparison with baselines in terms of (a) target enhancement; (b)~(c) interference suppression.**

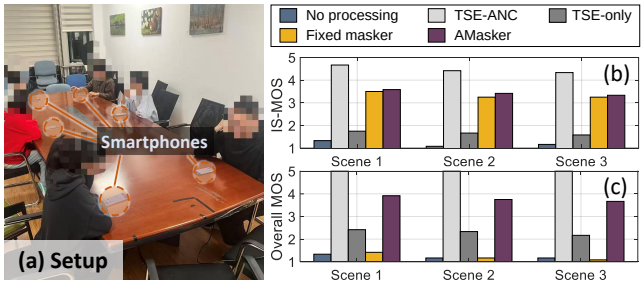



**Figure 14: Comparison with baselines in terms of human perception.**

*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.*

## 7.3 Overall performance

Here, we introduce an evaluation framework for intelligibility control capability and demonstrate that, even without ANC, 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 extraction with ANC (TSE-ANC)* [2]: For objective evaluation, we evaluate the target speech alone without interference, to simulate optimal noise-cancelling performance. For the subjective evaluation, we have volunteers wear AirPods Pro 3 and communicate via phone.

iii) *Target speech extraction (TSE-only)*: This method performs target speech extraction in the same way as AMasker, but only plays the extracted target speech amplified to 10 dB above interfering speech, without interference suppression.

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

[2]Note that TSE-ANC shares the same technical paradigm with current selective hearing works as introduced in Sec. 2.

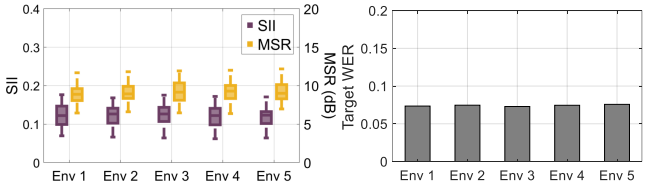



**Figure 15: Performance under different environments.**

**Comparison of target enhancement.** For each method, we test under different conversational scenes in Fig. 12 (a). Target WER results appear in Fig. 13 (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. For TSE-ANC, since interference is entirely eliminated, the average target WER drops below 0.03. However, ANC relies on headphones, limiting practical availability. 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 the unprocessed condition.

Although TSE-only and AMasker achieve similar target WERs, this does not render sound masking unnecessary. Low ASR WER does not mean that interfering speech is non-distracting to human listeners. This is because ASR models focus on the highest-SNR component without being drawn to intelligible interference, whereas human attention is easily captured by intelligible competing speech [64, 65]. Thus, merely evaluating target WER is insufficient; interference SII must also be measured to truthfully reflect performance.

**Comparison of interference suppression.** Figs. 13 (b) & (c) compare interference suppression of baseline methods via SII/MSR of interfering speech. The SII of the "No processing" method (0.7286 on average) and "TSE-only" method (0.5247 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 masker 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 invite volunteers to conduct conversations using different methods in Scene 3. 12 volunteers are randomly assigned to six two-person groups. For each method, two trials cover all participants. In each 10-minute trial, three

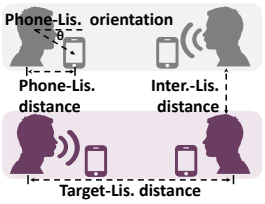



**Figure 16: Spatial parameters.**

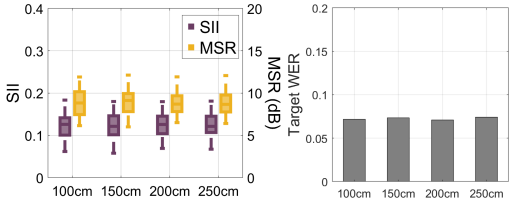



**Figure 17: Impact of target-listener distance.**

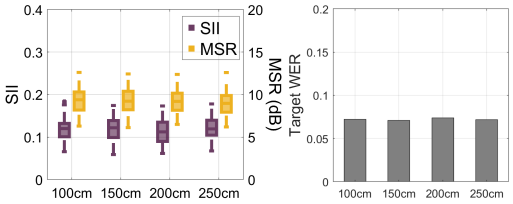



**Figure 18: Impact of interference-listener distance.**

groups hold concurrent conversations simultaneously. Afterward, each volunteer rates MOS as introduced in Sec. 7.2.3. Results appear in Fig. 14 (b) ~ (c). The TSE-only method has an average Interference Suppression MOS (IS-MOS) of 1.67, indicating most volunteers find interfering speech intelligible and distracting. In contrast, our method achieves an average IS-MOS of 3.44 and an average Overall MOS of 3.78, demonstrating noticeable improvement in both interference suppression and overall user experience. It validates the significance of selective sound masking. Note that the "Fixed masker" achieves similar IS-MOS to our method but lower Overall MOS as it also suppresses target speech.

**Remarks on potential perceptual artifacts.** It is worth noting that although the replayed and original target speech coexist for a listener, no volunteer reported perceptual disturbances or artifacts. This is because the total relay latency of target speech falls well below the 50 ms threshold of the *Precedence Effect* [66], a well-established psychoacoustic principle whereby two copies of the same sound arriving in close succession are perceptually fused into a single auditory event.

**Environmental generalization.** During training, data augmentation improves the model's robustness to variations in acoustic propagation, allowing AMasker to generalize well to the unseen basic environment shown in Fig. 12 (a). To further validate AMasker's generalization across different environments, we conducted experiments in 5 additional environments under the setting of Scene 2, as introduced in Sec. 7.2. The results in Fig. 15 show that the performance in Env. 1-5 is consistent with that in Sec. 7.3. Across all environments, the target speech WER stays below 0.08, while the SII and MSR remain below 0.2 and 13 dB, respectively.

## 7.4 Impact of spatial parameters

In this section, we evaluate AMasker under different spatial parameters shown in Fig. 16. All experiments are conducted in the primary environment shown in Fig. 12 (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.

*7.4.1 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. 17, 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.

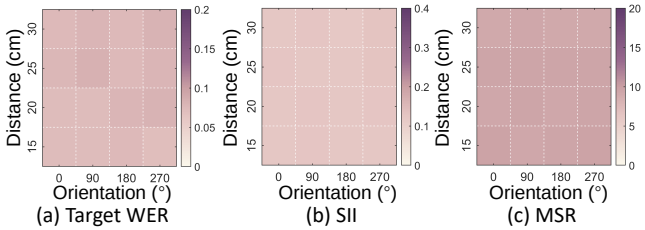



**Figure 19: Impact of phone-listener relative position.**

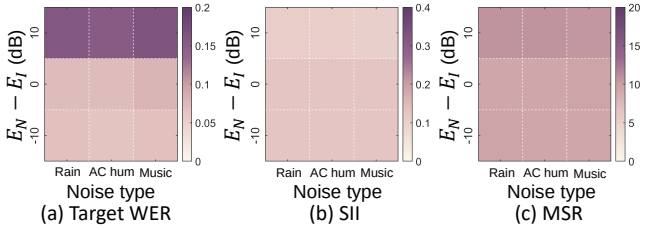



**Figure 20: Impact of non-intelligible noise.**

*7.4.2 Interference-listener distance.* We further test AMasker across distances between the interfering speaker and the listener. With a selective sound masking mechanism, AMasker can well reduce the undesired intelligibility without hurting the target intelligibility. We varied this distance from 1 m to 2.5 m, and the results in Fig. 18 show that AMasker still maintains stable performance with SII lower than 0.2, MSR lower than 15 dB, and WER lower than 0.1.

*7.4.3 Relative position between smartphone and listener.* In practice, users may casually place smartphones around them. The spatial offset between the smartphone and the user, on the one hand, slightly affects the SNR of target speech; on the other hand, as observed in Fig. 4 (c-2), it causes negligible spectral changes in the Bark scale, minimally affecting the masking efficacy. In this experiment, we vary both the orientation and distance between the smartphone and the listener, and test AMasker. The results in Fig. 19 validate AMasker's robustness under different relative positions, with WER, SII, and MSR remaining below the acceptable threshold.

## 7.5 Impact of practical factors

This section evaluates AMasker's performance under different practical factors. The environment and conversation setup are the same as Sec. 7.4.

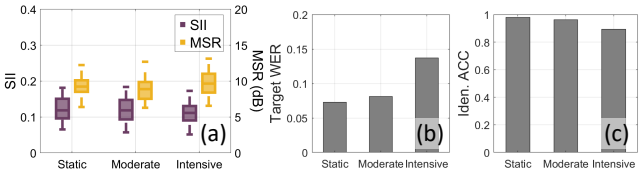



**Figure 21: Impact of mobility.**

*7.5.1 Non-intelligible noise.* AMasker targets scenarios where intelligible speech is the primary interference, while we also test it under different types and energy levels of unintelligible noise. Specifically, we played noise through a commercial speaker and, by calibrating with a sound pressure level meter, created different gaps between $E_N$ (unintelligible noise energy) and $E_I$ (interfering speech energy) at the listener. The results in Fig. 20 show that: i) When $E_N$ is lower than or comparable to $E_I$, AMasker remains unaffected. This is because, on the one hand, the SNR of target speech at the target speaker's smartphone remains high; on the other hand, the separation model can distinguish the intelligible interfering speech from the mixture and generate the specified masking sound. ii) When $E_N$ is significantly higher than $E_I$ (e.g., by 10 dB), noticeable performance degradation occurs, with the average WER of the target speech rising to 0.1587. This is because the high $E_N$ heavily reduces the input SNR of the target speech. Besides, SII is smaller since unintelligible noise also acts as a masking sound in masking performance assessment. Nevertheless, we believe such cases are rarely encountered in real-world scenarios.

*7.5.2 Mobility of users and devices.* In this experiment, we evaluate the performance of AMasker under different levels of mobility, including: *i) Static*, where participants and devices remain still; *ii) Moderate*, where we shake the dummy participant to simulate common movements of the human head (e.g., nodding), and move the device 50 cm away from the original place every 30 seconds; *iii) Intensive*, where we move both the participant and the device continuously within a 50 cm radius circle. The results in Fig. 21 (a)-(b) show that occasional device movement does not impair AMasker's performance, because when the identification model remains highly confident while the TDoA estimate is unstable, the TDoA label will be automatically updated within a few hundred milliseconds. However, under intensive mobility, the WER of the target speech exceeds 0.1, as highly volatile TDoA measurements force speaker identification to rely entirely on the lightweight DNN model, reducing accuracy below 90% (Fig. 21(c)). This causes portions of the target speech to be misclassified as interfering speech and masked.

## 7.6 Time and power consumption

We further evaluate the time and power consumption of AMasker's key modules on two representative smartphones: the Honor Magic 6 with Snapdragon 8 Gen 3 SoC and the iQOO 8 Pro with Snapdragon 888 SoC. We tested two types of modules: those requiring real-time online execution (e.g., speaker identification and speech separation) and those that can run offline (e.g., user embedding acquisition). For each module type, we executed it 100 times, and Table 2 lists its average latency and peak memory usage.

**Table 2: Overhead of AMasker's AI models.**

| Model | Type | Input (ms) | Latency (ms) | Memory usage (MB) |
|---|---|---|---|---|
| Iden (Magic 6) | online | 500 | 7.06 ± 0.50 | 74.21 |
| Iden (IQOO 8 Pro) | online | 500 | 11.18 ± 0.58 | 60.55 |
| Sep (Magic 6) | online | 12 | 8.56 ± 0.63 | 56.53 |
| Sep (IQOO 8 Pro) | online | 12 | 13.82 ± 0.74 | 49.69 |
| Emb (Magic 6) | offline | 5000 | 3465.71 ± 158.95 | 963.11 |
| Emb (IQOO 8 Pro) | offline | 5000 | 4778.83 ± 177.19 | 954.18 |

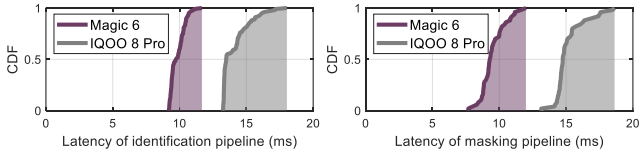



**Figure 22: Overall processing latency of real-time tasks.**

**Online modules.** The results in Table 2 demonstrate that both online AI models maintain ms-level latency even in a mid-range smartphone. Moreover, their combined memory usage is only ~130 MB and ~110 MB on Magic 6 and iQOO 8 Pro, respectively, which is acceptable given that current smartphones typically have more than 6 GB of RAM [67]. We then measure the overall processing latency for both the masking pipeline (speech separation and masking sound generation) and the identification pipeline. As shown in Fig. 22, the masking pipeline incurs average latencies of 9.52 ms and 15.18 ms on the Magic 6 and iQOO 8 Pro, respectively. Given the 32-ms total latency limit observed in Fig. 8, which leaves a 20-ms processing budget for the 12-ms window, our method consumes only 47.6% and 75.9% of this tolerance on the two devices, respectively. Furthermore, the identification pipeline has a maximum latency of only 18.03 ms on IQOO 8 Pro when processing a 500-ms window, ensuring rapid response to speaker transitions with a small hopping length.

**Offline module.** As shown in Table 2, the DNN-based user embedding acquisition module, which is not subject to real-time constraints, achieves an average latency of less than 4800 ms on both Magic 6 and iQOO 8 Pro, while requiring less than 1 GB of memory on each device. Note that the embedding module is only used for one-time voiceprint extraction. Once extracted, the voiceprint is reused across conversations and not recomputed repeatedly.

**Power consumption.** AMasker's whole-device power consumption, measured by Android's BatteryManager APIs, is 3.6 W and 3.3 W on Magic 6 and iQOO 8 Pro, respectively, both enabling>5 h of runtime on a 5000 mAh battery.

## 8 CONCLUSION

We present AMasker to enable Active Intelligibility Control (AIC) that leverages psychoacoustic insights: distraction primarily stems from speech intelligibility. AMasker employs a selective sound-masking workflow to enhance target intelligibility and precisely suppress interfering intelligibility. We implement a prototype and show its efficacy through extensive experiments.

## References


[1] Tobias Renz, Philip Leistner, and Andreas Liebl. Auditory distraction by speech: Can a babble masker restore working memory performance and subjective perception to baseline? *Applied Acoustics*, 2018.

[2] Densil Cabrera, Manuj Yadav, and Daniel Protheroe. Critical methodological assessment of the distraction distance used for evaluating room acoustic quality of open-plan offices. *Applied Acoustics*, 140:132–142, 2018.

[3] Marijke Keus van de Poll, Johannes Carlsson, John E. Marsh, Robert Ljung, Johan Odelius, Sabine J. Schlittmeier, Gunilla Sundin, and Patrik Sörqvist. Unmasking the effects of masking on performance: The potential of multiple-voice masking in the office environment. *The Journal of the Acoustical Society of America*, 138(2):807–816, 2015.

[4] Bandhav Veluri, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. Look once to hear: Target speech hearing with noisy examples. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2024.

[5] Tuochao Chen, Malek Itani, Sefik Emre Eskimez, Takuya Yoshioka, and Shyamnath Gollakota. Hearable devices with sound bubbles. *Nature Electronics*, pages 1–12, 2024.

[6] Sheng Shen, Nirupam Roy, Junfeng Guan, Haitham Hassanieh, and Romit Roy Choudhury. Mute: bringing iot to noise cancellation. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '18, page 282–296, New York, NY, USA, 2018. Association for Computing Machinery.

[7] S.M. Kuo and D.R. Morgan. Active noise control: a tutorial review. *Proceedings of the IEEE*, 87(6):943–973, 1999.

[8] MARKET RESEARCH FUTURE. Active noise cancellation headphones market research report. https://www.marketresearchfuture.com/reports/active-noise-cancellation-headphones-market-24552.

[9] MARKET RESEARCH FUTURE. Earphone and headphone market research report. https://www.marketresearchfuture.com/reports/earphone-headphone-market-7628.

[10] Marion Menanteau, Goël Fenech, Benjamin Adam, Eddy Langlois, Pierre Marcant, Eric Pelletier, Delphine Staumont-Sallé, Lynda Bensefa-Colas, and Marie-Noëlle Crepy. Severe allergic contact dermatitis from octylisothiazolinone in over-ear headphones: A case series. *Contact Dermatitis*, 92(4):291–298, 2025.

[11] Youyou Jin, Wei Hua, Qingfeng Liu, Zili Zheng, Xinyi Yao, Xiangling Zhang, Mei Li, Xiaoyun Zhang, Ran Gao, Xiaoyu Wang, et al. Otitis externa owing to allergic contact dermatitis to earphones and earplugs: A case control study. *JAAD International*, 2025.

[12] Annu Haapakangas, Valtteri Hongisto, Mervi Eerola, and Tuomas Kuusisto. Distraction distance and perceived disturbance by noise—an analysis of 21 open-plan offices. *The Journal of the Acoustical Society of America*, 141(1):127–136, 2017.

[13] Lisanne Bergefurt, Rianne Appel-Meulenbroek, and Theo Arentze. Level-adaptive sound masking in the open-plan office: How does it influence noise distraction, coping, and mental health? *Applied Acoustics*, 217:109845, 2024.

[14] Wolfgang Ellermeier and Jürgen 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*, 24(5):1406–1414, 1998.

[15] Joel Lewitz. Effective sound masking for speech privacy in open plan offices. *Journal of the Acoustical Society of America*, 123:2971, 2008.

[16] J. Keränen and V. Hongisto. Prediction of the spatial decay of speech in open-plan offices. *Applied Acoustics*, 74(12):1315–1325, 2013.

[17] L. Brocolini, E. Parizet, and P. Chevret. Effect of masking noise on cognitive performance and annoyance in open plan offices. *Applied Acoustics*, 114:44–55, 2016.

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

[19] Tobias Renz, Philip Leistner, and Andreas 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*, 143(3):EL212–EL217, 2018.

[20] Annu Haapakangas, Valtteri Hongisto, Jukka Hyönä, Joonas Kokko, and Jukka Keränen. Effects of unattended speech on performance and subjective distraction: The role of acoustic design in open-plan offices. *Applied Acoustics*, 86:1–16, 2014.

[21] Ankit 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.

[22] Sound masking systems. https://www.pureresonanceaudio.com/collections/sound-masking-systems. Accessed: 2025-01-28.

[23] Elite acoustics product line. https://www.elite-acoustics.com. Accessed: 2025-01-28.

[24] Valtteri Hongisto, David Oliva, and Laura Rekola. Subjective and objective rating of spectrally different pseudorandom noises—implications for speech masking design. *The Journal of the Acoustical Society of America*, 137(3):1344–1355, 2015.

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

[26] Lencore. Case study: Protecting speech privacy for cisco. https://www.lencore.com/case-studies/case-study-cisco/, n.d. Accessed: 2026-03-14.

[27] Joseph W Newbold, Jacob Luton, Anna L Cox, and Sandy JJ 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*, pages 2802–2809, 2017.

[28] Haworth. Acceptance & Efficacy of Biophilic Soundscaping in An Open-Plan Office. Research report, Haworth, 2021. Accessed on July 7, 2024.

[29] Jun Yang, Ming Wu, and Lu Han. A review of sound field control. *Applied Sciences*, 12(14):7319, 2022.

[30] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and Models*, volume 22. Springer Science & Business Media, 2013.

[31] Eberhard Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248–248, 1961.

[32] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[33] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, and Nancy L Dahlgren. Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1, 1993.

[34] Thomas H. Crystal and Arthur S. House. Segmental durations in connected-speech signals: Current results. *The Journal of the Acoustical Society of America*, 83(4):1553–1573, 1988.

[35] Caslav Pavlovic. Sii—speech intelligibility index standard: Ansi s3. 5 1997. *the Journal of the Acoustical Society of America*, 143(3_Supplement):1906–1906, 2018.

[36] Jennifer A Veitch, John S Bradley, Louise M Legault, Scott Norcross, and Jana 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.

[37] Yongjie Yang, Tao Chen, Zhenlin An, Shirui Cao, Xiaoran Fan, and Longfei Shangguan. LeakyFeeder: In-Air Gesture Control Through Leaky Acoustic Waves. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 144–157, UC Irvine Student Center. Irvine CA USA, May 2025. ACM.

[38] Karl S. Pearsons, Ricarda L. Bennett, and Sanford A. Fidell. Speech levels in various noise environments. Technical report, Office of Health and Ecological Effects, Office of Research and Development, US Environmental Protection Agency, 1977.

[39] Koenraad S Rhebergen and Niek J Versfeld. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 117(4):2181–2192, 2005.

[40] Brian C. J. Moore and Brian R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74(3):750–753, 1983.

[41] James D Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on selected areas in communications*, 6(2):314–323, 2002.

[42] Manfred R Schroeder, Bishnu S Atal, and JL Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America*, 66(6):1647–1652, 1979.

[43] Robert V Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304, 1995.

[44] Taishih Chi, Powen Ru, and Shihab A Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, 2005.

[45] Mounya Elhilali, Taishih Chi, and Shihab A Shamma. A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech communication*, 41(2-3):331–348, 2003.

[46] Michiko Kazama, Satoru Gotoh, Mikio Tohyama, and Tammo Houtgast. On the significance of phase in the short term fourier spectrum for speech intelligibility. *The Journal of the Acoustical Society of America*, 127(3):1432–1439, 2010.

[47] Hsin-Tien Chiang, Kuo-Hsuan Hung, Szu-Wei Fu, Heng-Cheng Kuo, Ming-Hsueh Tsai, and Yu Tsao. Study on the correlation between objective evaluations and subjective speech quality and intelligibility. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7, 2023.

[48] Ming Tu, Alan Wisler, Visar Berisha, and Julie M Liss. The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance. *The Journal of the Acoustical Society of America*, 140(5):EL416–EL422, 2016.

[49] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[50] Satwik Dutta, Shruthigna Chandupatla, and John Hansen. Adapting whisper for lightweight and efficient automatic speech recognition of children for on-device edge applications. *arXiv preprint arXiv:2507.14451*, 2025.

[51] Young-Jae Jang, Jaehyun Park, Won-Cheol Lee, and Hong-June Park. A convolution-neural-network feedforward active-noise-cancellation system on fpga for in-ear headphone. *Applied Sciences*, 12(11), 2022.

[52] Sheng Yue, Paul Pilon, and George Cavadias. Power of the mann–kendall and spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of hydrology*, 259(1-4):254–271, 2002.

[53] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[54] Pavel Zahorik. Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, 111(4):1832–1846, 2002.

[55] Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:136034, 2015.

[56] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier, 1978.

[57] Ragini Sinha, Ann-Christin Scherer, Simon Doclo, Christian Rollwage, and Jan Rennies. Evaluation of speaker-conditioned target speaker extraction algorithms for hearing-impaired listeners. *Trends in Hearing*, 29:23312165251365802, 2025.

[58] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. Tf-gridnet: Integrating full- and sub-band modeling for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3221–3236, 2023.

[59] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630, 2019.

[60] Zizheng Zhang, Chen Chen, Hsin-Hung Chen, Xiang Liu, Yuchen Hu, and Eng Siong Chng. Noise-Aware Speech Separation with Contrastive Learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1381–1385, Seoul, Korea, Republic of, April 2024. IEEE.

[61] Shikun Liu, Edward Johns, and Andrew 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.

[62] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, 1997.

[63] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

[64] Edward Collin Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25:975–979, 1953.

[65] Douglas S. Brungart. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3):1101–1109, 2001.

[66] Ruth Y Litovsky, H Steven Colburn, William A Yost, and Sandra J Guzman. The precedence effect. *The Journal of the Acoustical Society of America*, 106(4):1633–1654, 1999.

[67] Market.us Scoop. Android phone statistics 2025. https://scoop.market.us/android-phones-statistics/, 2025. Accessed: 2025-11-30.

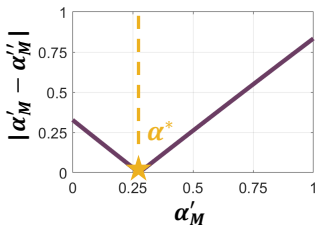



**Figure 23: Estimation error of flatness coefficient.**

# APPENDIX

## A Derivation of the ideal masking sound.

**Problem formulation.** Given a masking sound $M$, the masking threshold $TH_M$ can be estimated as:

$$TH_M = \psi(M) = (B_M * V_S - O(\alpha_M)) * V_S^{-1} \tag{6}$$

Here, $B_M$ is the Bark spectrum of masking sound, and $V_S$ is the spreading coefficient of the masking sound, which is a pre-defined vector that emulates how the critical bands of the masking sound interact with each other when the masking effect occurs [42]. $O(\alpha_M)$ is the offset between $TH_M$ and $B_M$ after spreading effect, which is determined by a flatness coefficient ($\alpha_M \in [0, 1]$) of the masking sound's power spectrum ($P(M)$) as:

$$O_i(\alpha_M) = (9 + i)\alpha_M + 5.5 \tag{7}$$

where $\alpha_M$ can be calculated when $P(M)$ is known as defined in [41]. Note that a more uniform energy distribution in $P(M)$ leads to a smaller $\alpha_M$ and a correspondingly smaller $O_i(\alpha_M)$, thereby reducing the required loudness of the masking sound when the masking threshold is known.

Given the interfering speech's Bark spectrum $B_I$, to achieve acceptable masking performance with minimal energy, an intuitive approach to obtain the ideal masking sound $M^*$ is to treat $B_I$ as the target masking threshold and solve the inverse problem $\psi^{-1}(B_I)$ to infer the power spectrum of $M^*$, denoted as $P^*(M)$. However, solving $\psi^{-1}(B_I)$ is challenging when the masking sound's power spectrum is unknown, since $\alpha^*_M$ is unknown and more than one ideal power spectrum $P^*(M)$ could lead to the same ideal Bark spectrum $B^*_M$ according to the Eq. 1, making the inversion an ill-posed problem.

**Solving the ill-posed inverse problem.** We solve the above problem based on the observation that the range of the flatness coefficient is finite and the ideal Bark spectrum $B^*_M$ can be derived when the optimal flatness coefficient $\alpha^*_M$ is determined. Specifically, by specifying a random flatness coefficient $\alpha'_M$, we could inversely calculate the corresponding Bark spectrum candidate $B'_M$. After that, we can guess a power spectrum candidate $P'(M)$ by uniformly distributing the power of each critical band in $B'_M$ across all frequency bins within that band. This uniform distribution is specifically designed to flatten the power spectrum as much as possible, thereby minimizing the offset $O_i(\alpha_M)$. Then, with existing equations, $P'(M)$'s corresponding flatness coefficient $\alpha''_M$ can be calculated. Since $\alpha'_M$ is a random guess, error may exist between $\alpha'_M$ and $\alpha''_M$. Apparently, $\alpha^*_M$ could be determined when the error $|\alpha'_M - \alpha''_M|$ is minimized. As shown in Fig. 23, the error of randomly guessed coefficients exhibits clear convex properties with only one global optimum when traversing $\alpha'_M$ from 0 to 1. Thus, we determine $\alpha^*_M$ by solving the objective function as:

$$\alpha^* = \arg\min_{\alpha'_M} |\alpha'_M - \alpha''_M| \tag{8}$$

We adopt a binary search to acquire $\alpha^*_M$ with an $O(\log(N))$ time complexity. The step of $\alpha'_M$ is set as 0.01 by default. With the inversion method mentioned above, $P^*(M)$ is determined after obtaining $\alpha^*_M$, and subsequently $M^*$ is determined by combining $P^*(M)$ and random phases.