MARCH 19 2025

Sparse representation of speech using an atomic speech model[a)]

Fanhui Kong; Huali Zhou; Nengheng Zheng; Qinglin Meng

J. Acoust. Soc. Am. 157, 1899–1911 (2025)
https://doi.org/10.1121/10.0036144

JASA ARTICLE

Sparse representation of speech using an atomic speech model[a)]

Fanhui Kong,[1,2] Huali Zhou,[3] Nengheng Zheng,[2] and Qinglin Meng[4,b)]

[1]*School of Information Engineering, Guangzhou Panyu Polytechnic, Guangzhou, Guangdong 410630, China*

[2]*Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong 518052, China*

[3]*School of Electronics and Information Engineering, Heyuan Polytechnic, Heyuan, Guangdong 517000, China*

[4]*Acoustics Laboratory, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, Guangdong 510641, China*

ABSTRACT:

Speech perception has been extensively studied using degradation algorithms such as channel vocoding, mosaic speech, and pointillistic speech. Here, an "atomic speech model" is introduced to generate unique sparse time-frequency patterns. It processes speech signals using a bank of bandpass filters, undersamples the signals, and reproduces each sample using a Gaussian-enveloped tone (a Gabor atom). To examine atomic speech intelligibility, adaptive speech reception thresholds (SRTs) are measured as a function of atom rate in normal-hearing listeners, investigating the effects of spectral maxima, binaural integration, and single echo. Experiment 1 showed atomic speech with 4 spectral maxima out of 32 bands remained intelligible even at a low rate under 80 atoms per second. Experiment 2 showed that when atoms were nonoverlappingly assigned to both ears, the mean SRT increased (i.e., worsened) compared to the monaural condition, where all atoms were assigned to one ear. Individual data revealed that a few listeners could integrate information from both ears, performing comparably to the monaural condition. Experiment 3 indicated higher mean SRT with a 100 ms echo delay than that with shorter delays (e.g., 50, 25, and 0 ms). These findings demonstrate the utility of the atomic speech model for investigating speech perception and its underlying mechanisms. © *2025 Acoustical Society of America*. https://doi.org/10.1121/10.0036144

(Received 26 August 2024; revised 19 February 2025; accepted 20 February 2025; published online 19 March 2025)

[Editor: Li Xu] Pages: 1899–1911

I. INTRODUCTION

In perceptual studies related to hearing, the use of degraded speech based on resynthesized acoustic stimuli remains a common method to deconstruct and analyze the underlying acoustic cues for intelligibility. Speech synthesized from a simplified hand-drawn spectrogram, played back via a pattern instrument, was indeed intelligible (Cooper *et al.*, 1952), providing a basis from which acoustic cues for phonemes were studied. Remez *et al.* (1981) demonstrated that speech intelligibility was preserved in the absence of traditional acoustic cues by tracking/replacing the first three formants of speech with sine waves. Temporal envelopes from a small number of channels can also offer high intelligibility (Shannon *et al.*, 1995; Dorman *et al.*, 1997), which, in turn, explains the rationale behind the earliest speaking machine (Dudley, 1939) and modern cochlear implants (CIs; Loizou, 2006). In recent years, the focus of research has shifted from speech in quiet environments to more complicated perceptual tasks (e.g., speech-in-noise recognition, pitch perception, and spatial hearing). Manipulations on temporal fine structure (Smith *et al.*, 2002) and spectral fine structure (Popham *et al.*, 2018) subsequently provided a more nuanced method of investigating the effects of speech degradation. As Gabor (1947, p. 591) put it, such methods "have already proved their heuristic value, and can be expected to throw more light on the theory of hearing." All of the evidence above confirms that under ideal listening conditions, the speech signal possesses significant redundancy, which can be quantitatively studied using these degradation algorithms and corresponding perception experiments.

This paper introduces an "atomic speech" synthesizer to further explore speech degradation. Natural sounds, including speech, are generally continuous in time. This work was inspired by recent studies in the field of CIs. When studying the temporal cues of speech, we analyze and manipulate the temporal envelope, periodicity, and fine structure, typically, by filtering in the continuous time domain. However, amplitude-modulated *discrete* electric pulse stimuli are used in modern CIs. To simulate the pulsatile electric hearing with CIs using acoustic stimuli in normal-hearing (NH) listeners, we recently proposed a Gaussian-enveloped tone (GET) vocoder (Meng *et al.*, 2023).

Proposed atomic speech model can be viewed as an extension of the GET vocoder model. The idea of the GET vocoder is to map each electric pulse into a Gaussian-enveloped acoustic tone burst. Although its performance is

[a)]Portions of this work have been included in the Ph.D. thesis submitted to Shenzhen University in 2023 by F.K.

[b)]Email: mengqinglin@scut.edu.cn

constrained by the time-frequency uncertainty principle, many key parameters of any conventional CI sound coding strategy can be simulated directly using the pulse-to-pulse mapping of the GET vocoder. In Meng *et al.* (2023), the GET vocoder was demonstrated by simulating the advanced combination encoder (ACE), a widely used CI strategy. When CI and NH listeners used comparable encoding parameters, they exhibited comparable speech perceptual patterns (Kong *et al.*, 2023). For CI simulation, the parameters of the GET vocoder were defined (or constrained) based on the parameters of ACE in our previous studies. For instance, the total number of channels was set to 22, the stimulation rate was set to 900 pulses per second (pps), and only the 8 highest peaks out of the 22 channels were preserved in each frame. Here, we extend the parameters of the GET vocoder to broader ranges, especially using much lower stimulation rates, allowing the model to generate new speech spectrogram patterns.

When the stimulation rate per channel is extremely low, such as 20–100 pps, the spectrogram of the synthesized speech exhibits sparsity in the temporal and spectral domains. As a result of the presence of numerous isolated GET atoms in the spectrogram, this type of speech is called atomic speech (Meng, 2020). A GET atom is characterized by Gaussian envelopes in the temporal and spectral domains, originally proposed by Gabor in the early 1940s as a fundamental component of sound (Gabor, 1947). In fact, similar concepts and terminology, "atomic music" or "microsound," have been employed in the field of music technology (Roads, 2004). These terms refer to the practice of working with extremely small sound units or atomic elements to create intricate and detailed musical compositions. Studies of atomic music or microsound focused on exploring the smallest components to achieve innovative and expressive outcomes in music.

Unlike previous research that focused on CI simulation with GET vocoders (Meng, 2020; Kong *et al.*, 2023), the primary objective of this study is to explore speech redundancy using the spectro-temporally sparse atomic speech. For studies with hearing-impaired individuals, natural speech waveform encoding is thought to be affected due to the loss of auditory nerve fibers. As a result, the speech waveform is considered stochastically undersampled, which significantly affects intelligibility, particularly in the cases of soft, rapid, and noisy speech (Lopez-Poveda and Barrios, 2013; Lopez-Poveda, 2014). In this study with NH listeners, atomic speech is also regarded as an undersampled speech waveform. However, it differs from the stochastic modeling of undersampling based on neurophysiology, as previously studied in the Lopex-Poveda's studies. Instead, atomic speech involves a deterministic undersampling of the speech waveform, achieved through a straightforward signal processing framework. This deterministic approach allows for a systematic and controlled exploration of speech redundancy, enabling a precise examination of the effects of spectro-temporal sparsity on speech intelligibility in NH listeners.

Reviewing recent literature on the sparsity and redundancy of speech signals reveals that several techniques share common interests with atomic speech, or in some cases, a similar conceptual framework. These include the glimpsing model (Cooke, 2006; Tang, 2022), ideal time-frequency segregation (Brungart *et al.*, 2006; Kjems *et al.*, 2009; Kidd *et al.*, 2016; Kidd *et al.*, 2019), sculpting speech (Cooke and Lecumberri, 2020), auditory bubble (Venezia *et al.*, 2016) and bubble noises (Mandel *et al.*, 2016; Mandel *et al.*, 2019), mosaic speech (Nakajima *et al.*, 2018; Santi *et al.*, 2020; Ueda *et al.*, 2022; Ueda *et al.*, 2024), and pointillistic speech (Kidd *et al.*, 2009). Among these techniques, the first three were used to examine energetic and informational masking based on local time-frequency signal-to-noise ratio (Brungart *et al.*, 2006; Cooke, 2006; Cooke and Lecumberri, 2020). The bubble-related methods aimed to measure spectro-temporal patterns (Venezia *et al.*, 2016) and the importance of time-frequency regions to intelligibility (Mandel *et al.*, 2016; Mandel *et al.*, 2019). Mosaic speech involved smearing the spectrogram in spectral and temporal domains (Nakajima *et al.*, 2018), unlike the channel vocoder, which mainly alters the spectral domain (Shannon *et al.*, 1995). Pointillistic speech reduced speech signals to a time-frequency matrix of brief pulse pure tones (Kidd *et al.*, 2009). Compared with these techniques, as will be demonstrated later, atomic speech exhibits distinctive spectro-temporal patterns and provides greater flexibility in controlling the pulsatile sparsity of a speech signal.

Among these techniques, mosaic speech, pointillistic speech, and atomic speech all rely on a time-frequency matrix as their foundation, decomposing speech into grids of time and frequency components. Whereas mosaic speech reconstructs signals using larger acoustic patches, pointillistic speech uses discrete sound "points" for a minimalist effect, and atomic speech focuses on sparse, essential elements that preserve critical auditory cues. These distinctions highlight different strategies for manipulating and understanding speech while sharing a common analytical framework. Figure 1 illustrates the processing results of the three techniques applied to the same original signal. In mosaic speech, the energy of the original speech within each grid cell is averaged and replaced by noise carriers within the same grid. In pointillistic speech, the energy of the original speech within each grid cell is similarly averaged but expressed using short sinusoidal carriers with a fixed duration. In atomic speech, the time-frequency point with the highest energy within each grid cell is replaced by a shorter pure tone, such as a GET (or Gabor atom), and is further refined by comparing time-frequency points across grids within the same time frame to retain only those with the highest amplitude. Although atomic speech and pointillistic speech break down speech into a series of pulsed pure tones, atomic speech achieves greater sparsity in the time-frequency plane by using short Gabor atoms and employing a much larger number of channels. In contrast, pointillistic speech remains continuous in the time domain (Kidd *et al.*, 2009).

https://doi.org/10.1121/10.0036144

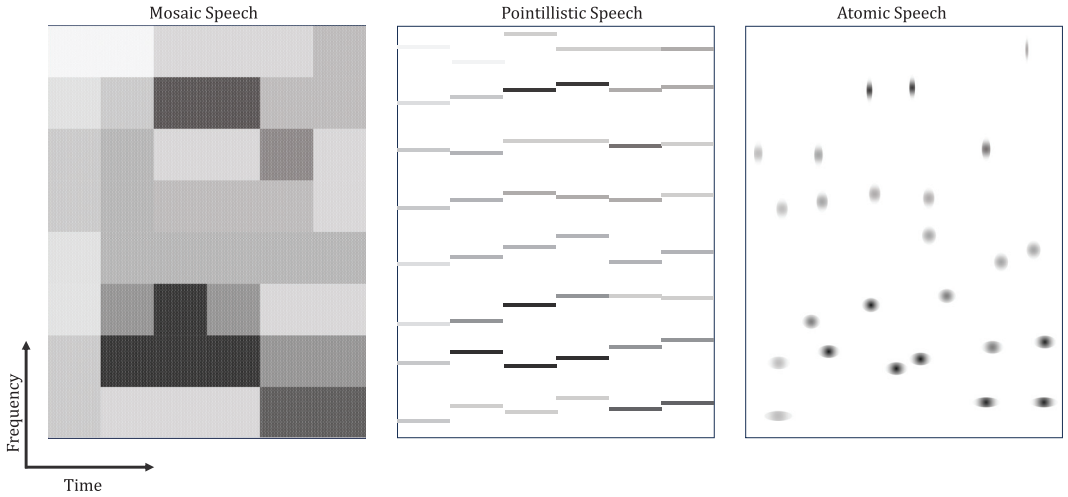

FIG. 1. Schematic illusions of the mosaic speech (8 channels), pointillistic speech (8 channels), and atomic speech (4 maxima selected from 16 channels), which are all derived from the same original speech signal.

Specifically, the atomic speech model undersamples the band-passed signals from N (e.g., 32) channels at M ($M \leq N$) spectral peaks (or maxima) among within-frame temporal peaks using Gabor atoms. In this procedure, the peak points in the time-frequency plane were assumed to deliver the most information among all time-frequency points. The frame duration could be used to control the atom rate and, consequently, influence, together with the spectral maxima number, the sparsity of the generated atomic speech.

Neither the sparse patterns nor the intelligibility of atomic speech has been well documented in the literature. As mentioned above, different models study sparsity or redundancy by reducing cues or degrading information from the signals according to different criteria, e.g., signal-to-noise threshold for speech-on-speech condition (Brungart *et al.*, 2006; Cooke, 2006; Cooke and Lecumberri, 2020) and temporal and/or spectral smearing degree for clean speech (Shannon *et al.*, 1995; Nakajima *et al.*, 2018). The corresponding criterion for the atomic speech model is the sparsity, which is controlled by the atom rate and/or the number of spectral maxima.

To study the effects of sparsity on intelligibility for clean speech using atomic speech, in our first experiment, we synthesized atomic speech with varying stimulation rates (or atom rates) and the number of spectral maxima. Based on our specific implementation of the model, the lowest necessary rate for 50% speech understanding was measured using an adaptive speech reception threshold (SRT) procedure. To demonstrate the flexibility of manipulating atomic speech parameters, we applied dichotic processing in SRT experiment 2 to investigate binaural integration. In SRT experiment 3, we introduced a single echo into the model to examine the effects of single reflections on speech intelligibility. For experiment 2, considering the variability in binaural integration observed in prior studies using some types of dichotic speech stimuli (e.g., Spehar *et al.*, 2008), we hypothesized that while NH listeners can integrate dichotic acoustic cues, randomly assigning speech "atoms" to different ears would hinder speech understanding compared to presenting the same information to one ear, despite individual variability in binaural integration capabilities. For experiment 3, given the established influence of early and late reflections on speech intelligibility (Warzybok *et al.*, 2013), we aim to investigate how single echoes with varying delays interact with the sparse spectro-temporal structure of atomic speech, offering a new perspective on echo perception.

Therefore, the Secs. II and III introduce the signal processing of the proposed model and the three experiments in NH listeners to demonstrate the effects of the number of spectral maxima, binaural integration, and single echos on the intelligibility of atomic speech. Then, the findings from the experiments and their relationship to existing models and related findings are discussed.

II. ATOMIC SPEECH MODEL

A. Proposed signal processing framework

The implementation of the atomic speech model relies on a bank of bandpass filters. A consensus is widely recognized that individual points on the basilar membrane function like bandpass filters. Auditory modeling and applications rooted in bandpass filters are also frequently encountered. In the atomic speech model, it is postulated that each output from these filters comprises numerous Gabor atoms (Gabor, 1947), which exhibit compactness in the temporal and spectral domains. At its core, this model aims to depict each output from the bandpass filters as a connection of Gabor atoms, enabling the manipulation of speech for diverse objectives through atom modifications.

The atomic speech model encompasses a series of stages aimed at attaining its distinctive sparsity in the spectro-temporal domain. These stages encompass the subsequent components:

(1) *Band-pass filtering*: The input speech is passed through a bank of band-pass filters to obtain frequency-specific signals;
(2) *frame division*: The band-passed signals are segmented into frames, which allows for localized analysis of the temporal properties;
(3) *temporal peaks selection*: In each frame of the temporal envelope for each frequency channel, temporal peaks are identified;
(4) *spectral maxima selection*: Among the temporal peaks identified in each frame, the spectral maxima are chosen, indicating the most dominant frequencies at those time instances;
(5) *GET synthesis*: At each selected temporal and spectral peak, a GET is synthesized, representing a compact spectro-temporal atom; and
(6) *playback*: Finally, the synthesized GETs are played back at the corresponding times and frequencies, generating the atomic speech signal.

B. Specific implementation used in our experiments

Section II A gives the general conceptual framework of the atomic speech model, and all stages can be flexibly manipulated using arbitrary signal processing methods according to research questions. The specific implementation of the atomic speech model used in our experiments is depicted in Fig. 2.

The MATLAB code (The MathWorks, Natick, MA) for this implementation is accessible on GitHub.[1] The speech signal $s[n]$, sampled at 16 kHz, underwent an initial decomposition into 32 channels using a bank of 2-pass sixth-order Butterworth filters defined by the butter.m function in MATLAB. The filtering process was accomplished using zero-phase filtering, implemented with the filtfilt.m function in MATLAB. The cutoff frequencies for the filters were defined using the equivalent rectangular bandwidths (ERBs) scale of the human auditory system (Moore, 2013). These cutoff frequencies were defined to cover the frequency range from 80 to 7990 Hz. The Hz to ERB scale conversion is defined as

$$\mathrm{ERB} = 24.7 \times \left(4.37 \times 10^{-3} \times f + 1\right), \tag{1}$$

where f is the frequency in Hz. The corresponding ERB to Hz scale conversion was used to compute the cutoff frequencies of the bandpass filters, resulting in the following cutoff frequencies for the 32 channels: 80, 113, 150, 191, 237, 287, 343, 404, 473, 548, 632, 725, 828, 943, 1069, 1209, 1364, 1536, 1727, 1938, 2172, 2432, 2719, 3037, 3390, 3781, 4214, 4693, 5225, 5814, 6466, 7189, and 7990 Hz. The output signal from the i th bandpass filter was denoted as $x_i[n]$.

The temporal envelope of each channel was obtained by applying the Hilbert transform and full-wave rectification to the filtered signals $x_i[n]$. It was further smoothed using a third-order Butterworth low-pass filter (LPF). The smoothed envelope $e_i[n]$ for the i th channel is computed as

$$e_i[n] = \mathrm{LPF}\Big(\big|\mathrm{Hilbert}(x_i[n])\big|\Big), \tag{2}$$

where LPF was performed using the butter.m and filtfilt.m functions in MATLAB. The $|.|$ operation denotes the modulus or absolute value of a complex number, and Hilbert represents the Hilbert transform or the calculation of the analytic signal, performed using the hilbert.m function in MATLAB. The cutoff frequency of the LPF, denoted as $f_{c,\mathrm{LPF}}$, was determined by the rule

$$f_{c,\mathrm{LPF}} = \min(200, 0.5 \times r_{\mathrm{stim}}), \tag{3}$$

where r_{stim} represents the stimulation rate per channel in Hz (or pps). This parameter can be user-defined or software-controlled and is used in the subsequent steps of the atomic speech model. When r_{stim} is higher than 400 Hz, the LPF will have a fixed cutoff frequency of 200 Hz. However, if

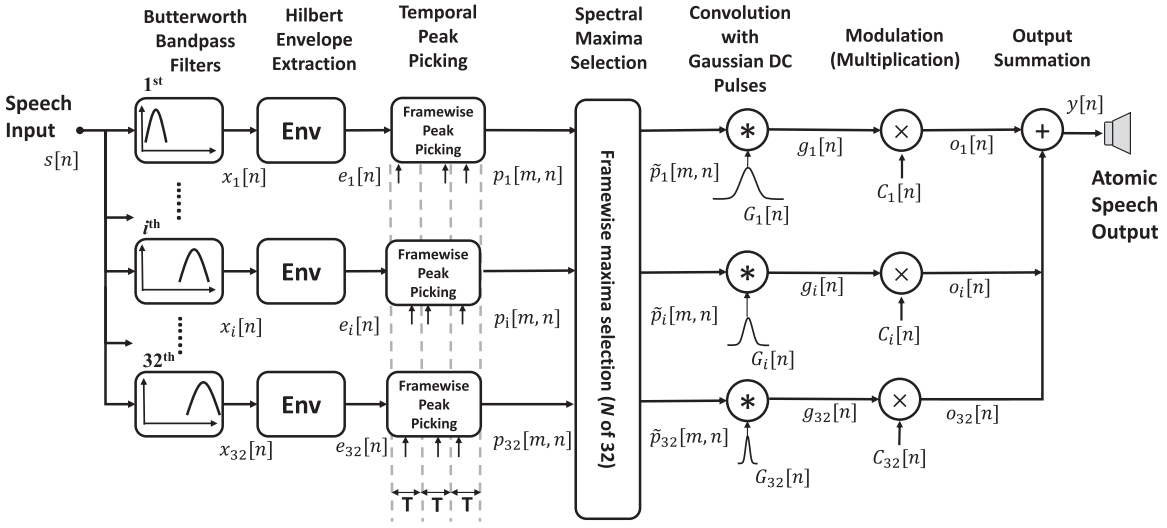

FIG. 2. A specific implementation flow chart of the atomic speech model.

https://doi.org/10.1121/10.0036144

the stimulation rate is lower than or equal to 400 Hz, the cut-off frequency will be half of the stimulation rate (0.5 times r_{stim}) to ensure a reasonable level of smoothing.

Then, the envelope $e_i[n]$ was divided into frames with a sample size of $K = \lfloor f_s/r_{\text{stim}} \rceil$. The $\lfloor . \rceil$ denotes rounding to the nearest integer. The frameshift was also set to K. The frame duration $T = K/f_s$. There is no overlap between frames. In the mth frame, denoted by $p_i[m,n]$, the temporal peak was preserved and the other samples were set to zero. The temporal peak was defined as the sample with the maximum value in the frame.

In each frame, the amplitudes of the temporal peaks among all 32 channels were compared. Among these peaks, only the highest N peaks were preserved, and the other peaks were set to zero. The parameter N was defined as the number of spectral maxima, later denoted by *max_Ch* in Secs. III A–III C (the experiments). This step is similar to the spectral maxima selection in the ACE CI strategy (Vandali *et al.*, 2000; Kong *et al.*, 2023). The new frame signal for the i th channel was denoted as $\tilde{p}_i[m,n]$. By connecting all of the frames for the i th channel, the new temporal-spectrally sparse signal was denoted as $\tilde{p}_i[n]$. It can be expressed as the sum of individual frame signals $\tilde{p}_i[m,n]$ such that

$$\tilde{p}_i[n] = \sum_{m=1}^{M} \tilde{p}_i[m,n] = \sum_{m=1}^{M} A_i[m]\delta[n-(mK-k_{i,m})], \tag{4}$$

where M is the total number of frames, $A_i[m]$ is the amplitude of the peak in the m th frame, and $k_{i,m} \in [0,K]$ is the actual time shift of the peak in the m th frame from the frame end. If $k_{i,m} = 0$, it means that the peak in the i th channel is located at the frame end. On the other hand, if $k_{i,m} = K$, it means that the peak is positioned at the frame start. If the peak of the i th channel was not selected in the m th frame, $A_i[m] = 0$.

Then, $\tilde{p}_i[n]$ is convolved with a Gaussian window function $G_i[n]$, denoted as $g_i[n]$, using the convolution operation "$*$" such that

$$g_i[n] = \tilde{p}_i[n] * G_i[n] = \sum_{m=1}^{M} A_i[m]G_i[n-(mK-k_{i,m})]. \tag{5}$$

The Gaussian window function $G_i[n]$ is defined as

$$G_i[n] = \frac{1}{\sqrt{D_i}} \exp\left(-\pi\left(\frac{n}{f_s}\right)^2 \Big/ D_i^2\right), \tag{6}$$

where D_i represents the effective duration of the Gaussian window for the i th channel and is defined as the reciprocal of 1.019 times the ERB

$$D_i = \frac{1}{b} = \frac{1}{1.019 \times \text{ERB}} = \frac{1}{1.019 \times 24.7 \times \left(4.37 \times f_i^c/1000 + 1\right)}. \tag{7}$$

In Eq. (7), f_i^c is the center frequency of the i th channel in Hz. The factor 1.019 has been used in many auditory filter modeling studies (Patterson *et al.*, 1992; Patterson, 2000). The output signal for each channel $o_i[n]$ is obtained by point-wise multiplying the convolution output with a sine wave of frequency f_i^c and a randomly distributed initial phase ϕ_i in the range $[0, 2\pi]$ such that

$$o_i[n] = g_i[n] \times C_i[n] = g_i[n] \times \sin\left(\frac{2\pi f_i^c n}{f_s} + \phi_i\right). \tag{8}$$

The final output of the atomic speech $y[n]$ is obtained by summing up all of the channels

$$y[n] = \sum_{i=1}^{32} o_i[n]. \tag{9}$$

To ensure a consistent loudness level, the level of the final output atomic speech $y[n]$ is root-mean-square (rms)-normalized to be the same as the original speech signal $s[n]$. Finally, the normalized atomic speech is delivered to an audio interface and headphones in a sound-proof room for evaluation.

The synthesis process described above completes the generation of atomic speech, which is an innovative speech redundancy manipulation algorithm based on spectro-temporally sparse atomic patterns. Figure 3 demonstrates an atomic speech with an atom rate of 100 pps and a maxima number of one. The highly sparse speech is still possible to deliver most word meaning to NH listeners as indicated in the experiments. Three experiments in perception of speech in quiet were conducted to showcase the recognition of atomic speech. These experiments individually investigated the impacts of the number of spectral maxima, binaural integration, and echo delays.

III. SPEECH PERCEPTION EXPERIMENTS

A. Experiment 1: Spectral maxima number

The primary objective of the first experiment was to determine the intelligibility of atomic speech across varying spectral maxima numbers. Additionally, the study examined the trade-off between temporal and spectral resolution by adjusting the atom rate (temporal aspect) and number of maxima (spectral aspect) in the atomic speech model. Similar investigations into this trade-off have been conducted by Xu and colleagues using channel vocoders, where they manipulated the number of channels and cutoff frequencies of temporal envelopes within each channel (Xu *et al.*, 2002; Xu *et al.*, 2005; Xu and Zheng, 2007; Xu and Pfingst, 2008).

Ten naive college students (five were females and aged between 19 and 32 years old) with normal hearing were selected for the study. All participants were native Mandarin speakers. Prior to the experiment, the participants underwent a screening for normal hearing using a validated automated

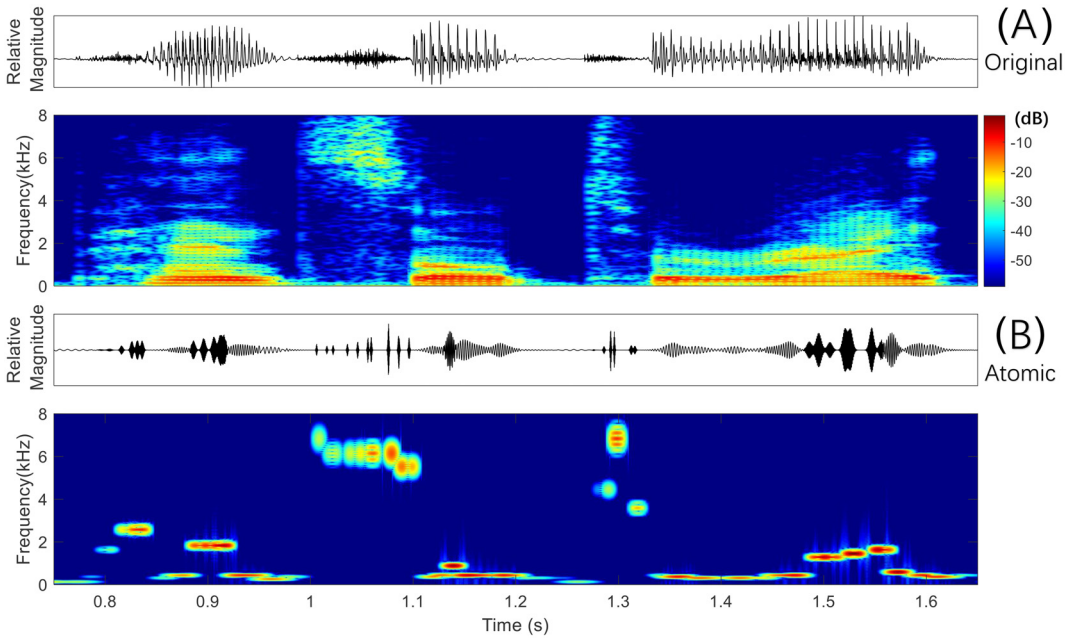

FIG. 3. A demonstration of the atomic speech. (A) An original English speech "pencil to write" and (B) an atomic speech generated using the specific settings of the atomic speech model as described in this section are shown, with an atom rate of 100 pps and a maxima number of one.

audiometer (Guo *et al.*, 2021). Their hearing thresholds were measured at octave frequencies ranging from 250 to 8000 Hz, and all thresholds were found to be 20 dB hearing level (HL).

To ensure compliance and adherence to ethical guidelines, written consent forms were obtained from all participants before the commencement of the experiment, as approved by the ethical review committee at Shenzhen University. Additionally, participants were compensated for their time and contribution to the study. These procedures were followed in experiments 2 and 3 as well.

For conducting the experiment, the Mandarin Hearing in Noise Test (MHINT) corpus (Wong *et al.*, 2007) was used. The corpus was produced by a male speaker and is well-established as a standard resource for speech perception studies in Mandarin (Xu *et al.*, 2021). The corpus comprises a total of 14 lists, including 12 lists specifically designed for testing and 2 lists intended for training purposes. Each testing list is composed of 20 sentences with each sentence containing 10 words. Participants were instructed to repeat the words in each sentence. The speech stimuli were presented diotically through high-quality headphones at a comfortable volume level.

The SRT was determined as the rate of atoms (or pulses) per channel, measured in pps, which resulted in 50% intelligibility for the atomic speech. This measurement was obtained using a one-down one-up adaptive procedure, where the atom rate was adjusted based on the participant's responses to achieve the 50% intelligibility threshold. The initial rate for the adaptive tracks was set to achieve a total rate of 800 pps across all channels, which showed high intelligibility in pilot experiments. Specifically, the adaptive tracks for each condition started with an initial rate of $800/max\_Ch$ pps (i.e., 800, 200, and 100 pps for $max\_Ch = 1$, 4, and 8, respectively). The rate was adjusted based on positive or negative responses. The rate change factors were 0.5 before the second reversal, 0.25 before the fourth reversal, and 0.125 after the fourth reversal. Calculated rates were rounded up to the nearest integer. The SRT was determined by averaging the last eight rates in the adaptive track.

Participants were tested under three conditions ($max\_Ch = 1$, 4, and 8). Each condition was evaluated in three separate runs with different sentence lists. The first run was a training session to familiarize participants with atomic sounds and the procedure. The final result for each condition was the average of the SRTs from the last two runs. The order of maxima numbers and lists was randomized across participants to minimize biases. In the training run, feedback was given, and participants could repeat stimuli as desired. In testing runs, no feedback was provided, and stimuli were presented only once.

SRT results of rate per channel and total rate for all channels are shown in Figs. 4(E) and 4(F), respectively. Statistical analysis was performed using a one-way repeated measures analysis of variance (rm-ANOVA) using Greenhouse-Geisser corrections for sphericity. Results indicated a significant main effect of maxima number for Fig. 4(E) [$F(1.205, 10.84) = 15.49, p = 0.002$] and Fig. 4(F) [$F(1.008, 9.070) = 124.0, p < 0.0001$]. *Post hoc* analysis using Tukey's honestly significant difference (HSD) test was used for multiple comparisons. The mean SRTs were significantly lower with larger max_Ch ($p < 0.05$) and 105.6, 18.2, and 12.6 pps per channel, respectively, for $max\_Ch = 1$, 4, and 8. For the total rate for all channels, the difference between $max\_Ch = 1$ and 8 was insignificant ($p > 0.05$), but the mean SRTs with $max\_Ch = 4$ (72.7 pps)

https://doi.org/10.1121/10.0036144

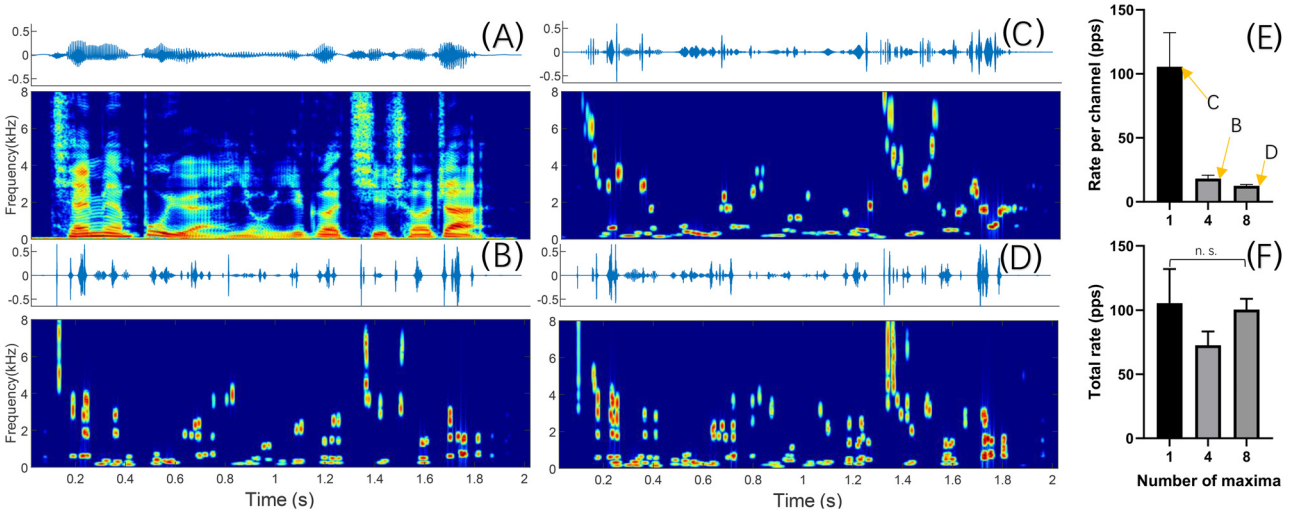

FIG. 4. SRT results (mean + SD) with atomic speech using three numbers of maxima (one, four, and eight) are shown for rate per channel (E) and total rate for all channels (F). Demonstrations of the atomic speech signals [(B)–(D)] generated using the proposed framework and program, with (A) representing the original signal. The rates for (B)–(D) are the mean SRTs for the three conditions as indicated by the arrows in (E).

were significantly lower than those with $max\_Ch = 1$ (105.6 pps) and 8 (100.6 pps; $p < 0.05$).

To our knowledge, the intelligible degraded patterns of atomic speech depicted in Fig. 4 have not been previously reported in the existing literature. Notably, for the four-maxima condition, an average total atom rate of 72.7 pps was sufficient for participants to understand the content of one sentence, despite the speech being highly sparse in spectral and temporal domains. This implies that speech understanding does not depend on highly sampled speech signals. In CIs, the default rate is usually higher than 900 pps multiplied by more than eight channels, i.e., a minimum of 7200 pps. In auditory nerve modeling studies (Lopez-Poveda and Barrios, 2013; Lopez-Poveda, 2014), it has been demonstrated that stochastic undersampling degrades speech intelligibility in noise rather than in quiet. Atomic and stochastic modeling studies suggest that the auditory system may not need highly sampled speech signals for speech understanding. This is also one aspect of speech redundancy.

The SRT for the total rate with $max\_Ch = 4$ was lower compared to those with $max\_Ch = 1$ and 8 maxima. In theory, this might be attributed to the presence of more independent atoms when $max\_Ch = 4$. Although we assumed that the atoms in atomic speech are isolated in the time-frequency plane, the model algorithm does not guarantee this isolation. With the same total rate, the $max\_Ch = 1$ condition may lead to more temporal interaction because of higher within-channel rates. For example, in Fig. 4(C), many atoms at frequencies lower than 1 kHz are connected with their temporal neighbors, forming horizontal lines. On the other hand, the $max\_Ch = 8$ condition may lead to more spectral interaction as a result of the high number of maxima. For instance, in Fig. 4(D), many atoms at frequencies higher than 4 kHz are connected with their spectral neighbors, resulting in vertical lines. Consequently, under the conditions of one and eight maxima, there was increased inter-atom interference or decreased atom independence. The $max\_Ch = 4$ condition, as illustrated in Fig. 4(B), exhibits reduced temporal interaction compared to the $max\_Ch = 1$ condition and decreased spectral interaction compared to the $max\_Ch = 8$ condition. Moreover, the comparable mean SRTs between the one and eight maxima conditions indicated a trade-off between the spectral and temporal domains, which is consistent with previous works using channel vocoders (Xu *et al.*, 2005; Xu and Pfingst, 2008). In the subsequent experiments, we chose to use four maxima for atomic speech generation, given its potential benefits in achieving better speech intelligibility and reduced inter-atom interference.

B. Experiment 2: Binaural integration

In the second experiment, we explored the potential of atomic speech for binaural hearing studies. Specifically, we compared a dichotic condition with a monaural condition. This comparison aimed to investigate the effects of binaural integration on speech perception when using atomic speech. A detailed discussion of the relationship between this experiment and existing literature on binaural integration is provided in Sec. IV.

In both conditions, the stimuli were generated based on the four-maxima condition from experiment 1. For the monaural condition, the atomic speech was presented to the left ear only. In contrast, the dichotic condition involved randomly assigning the atoms to the left and right ears in a balanced manner, where half of the atoms are presented to each ear. Therefore, both conditions shared the same model parameters and the same original speech signal, meaning they contained the same "information." The key difference between the two conditions was the presentation of the information. The monaural condition delivered all of the information to the left ear while the dichotic condition

distributed the information half-by-half to both ears, creating a binaural listening experience.

We hypothesized that presenting the same information in the dichotic condition would be more challenging to understand compared to that for the monaural condition. This is because the dichotic condition necessitates the listener to integrate the information from both ears to comprehend the speech, whereas the monaural condition does not require such integration. Another reason why one might expect that binaural integration could be hard is that the introduced binaural cues (very large interaural level differences for different time-frequency regions) are rarely experienced in real-life scenarios. To test this hypothesis, we compared the SRTs for the two conditions. The study involved 20 naive NH college students (12 females, aged 19–25 years old) who were native Mandarin speakers. The testing procedure was identical to that used in experiment 1.

SRT results of rate per channel for the monaural and dichotic conditions are depicted in Figs. 5(E) and 5(F). The mean SRT was 19.4 pps/ch for the monaural condition and 23.0 pps/ch for the dichotic condition. The corresponding values for the total rate of all channels are 77.8 and 92.0 pps, respectively. The mean SRT for the monaural condition (19.4 pps/ch or 77.8 pps total) was similar to the diotic condition in experiment 1 (18.2 pps/ch or 72.7 pps total). Statistical analysis was conducted between the monaural and dichotic conditions using a paired-sample t-test. Results showed a significant difference between the two conditions $[t(19) = 4.009, p < 0.001]$. The mean difference was 3.6 pps/ch [see Fig. 5(E)] or 14.2 pps in total. These findings supported our hypothesis that the dichotic condition was more challenging than the monaural condition in terms of speech perception using atomic speech.

Although the group means showed a significant difference between the two conditions, individual results [Fig. 5(F)] revealed that the dichotic condition was not consistently more challenging than the monaural condition for all participants. Among the 20 participants, 5 individuals exhibited slightly better results in the dichotic condition while 10 participants experienced much worse performance in the dichotic condition with a difference larger than 3.0 pps/ch. The variance in the dichotic condition was larger than that in the monaural condition, possibly, indicating that the integration of information from both ears was more difficult for some listeners than for others. This variability could be attributed to individual differences in binaural processing abilities and auditory integration skills.

In the dichotic condition, the stimuli presented to the two ears had nonoverlapping two-dimensional Gaussian envelopes in the temporal and spectral domains. This was because the atoms were randomly assigned to the two ears in a balanced manner. As a result, the atoms had different directional cues between the left and right ears. The listening task in the dichotic condition, especially around the threshold, involved integrating the atoms with distinct directions (i.e., left or right) and nonoverlapping spectrotemporal envelopes to comprehend the speech. This integration task could be more challenging than the monaural condition, where all Gaussian atoms had the same directional cues (i.e., from the same ear side). However, despite the potential difficulty, the results showed that some listeners were able to effectively integrate information from both ears and perform better than or comparably with the monaural condition. This highlights the individual variability in binaural processing abilities and the capacity to extract meaningful speech information from stimuli with nonoverlapping Gaussian envelopes. These findings contribute to a better understanding of binaural hearing and the potential benefits of using atomic speech for studying binaural integration in speech perception.

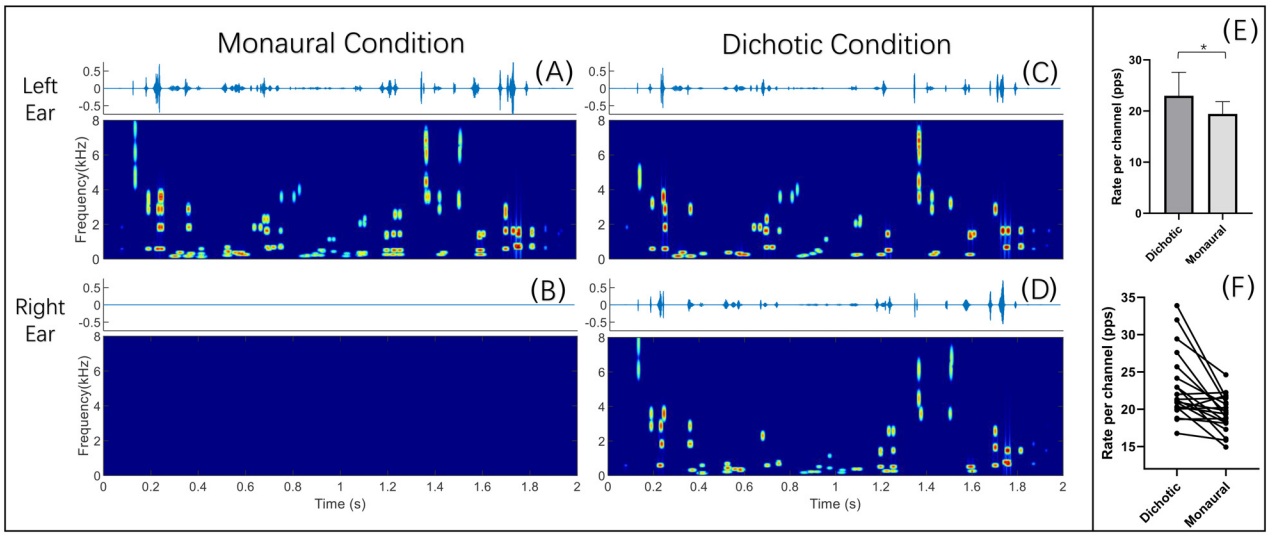

FIG. 5. SRT results [rate per channel in pps, (E) mean + SD with an asterisk, indicating a significant difference and (F) individual results] with atomic speech using four maxima with the monaural and dichotic conditions and demonstrations of the two atomic speech signals [(top) left ear sounds and (bottom) right ear sounds)].

https://doi.org/10.1121/10.0036144

C. Experiment 3: Single echo

Experiment 1 measured the SRTs with diotic atomic speech under three different spectral maxima, i.e., one, four, and eight. Experiment 2 measured the SRTs with dichotic and monaural atomic speech under four spectral maxima. In experiment 3, we measured the SRTs with diotic atomic speech under four spectral maxima with four different echo delays, i.e., 0, 25, 50, and 100 ms. The echo delay refers to the time difference between the direct and reflected signals, which can have a significant impact on speech perception in reverberant environments.

The echo signal was implemented by delaying the direct signal by the corresponding time. To create the final stimulus, the direct and echo signals were summed together. Previous studies have demonstrated that early reflections, up to 50 ms, can improve the intelligibility of the direct speech by increasing loudness, whereas later reflections may have a detrimental effect on intelligibility (Li *et al.*, 2015). In our implementation, the loudness cue was mostly unfeasible because of the normalization. We hypothesized that the late echo with a 100 ms delay would also negatively influence speech understanding with atomic speech. A detailed discussion of the relationship between this experiment and existing literature on reflection effects in NH listeners is provided in Sec. IV.

SRTs for the four delays were measured. Ten naive NH college students (five females; aged 19–25 years old; native Mandarin speakers) were tested. The testing procedure was the same as that in experiment 1. The stimuli were presented diotically.

The results are depicted in Figs. 6(E) and 5(F). The mean SRTs were 18.2, 16.3, 18.3, and 33.5 pps/ch for the 0, 25, 50, and 100 ms delay conditions, respectively. The corresponding values for the total rate of all channels were 72.8, 65.2, 73.2, and 134.0 pps. Statistical analysis was conducted between the four delay conditions using a one-way rm-ANOVA. Results showed a significant effect [$F(1.437, 12.93) = 0.027$, $p < 0.001$]. *Post hoc* analysis using Dunnett's test showed that the 100 ms delay condition was significantly different from the no echo (0 ms) condition ($p = 0.006$), and the 25 and 50 ms conditions were not significantly different from the no echo condition ($p = 0.331$ and $p > 0.999$, respectively).

The results indicated that the echo delay of 100 ms significantly worsened speech intelligibility (increased the SRTs). This holdup caused the delayed atoms from a leading phoneme to potentially interact with the atoms from a following phoneme and fill in the inter-phoneme gaps. These effects may have had a detrimental impact on speech understanding. In contrast, for the shorter delays, the reflected atoms did not seem to harm speech understanding to the same extent. Consequently, our hypothesis regarding the effect of echo delay on speech perception with atomic speech was supported by the experimental results.

IV. DISCUSSION

In this section, we summarize the major findings from our study. First, we examine the SRTs obtained with atomic speech, investigating the effects of various physical parameters such as atom rate, spectral maxima, binaural cues, and echo delay. Through three adaptive SRT experiments, we demonstrate how atomic speech can offer insights into fundamental aspects of speech perception.

A. SRTs of atomic speech

Experiment 1 demonstrated that presenting salient atoms in the spectro-temporal domain could suffice for speech comprehension, even without explicit representation of formants and harmonics. Consistent with earlier investigations into degraded speech, such as sine-wave speech (Remez *et al.*, 1981; Feng *et al.*, 2012) and vocoded speech (Shannon *et al.*, 1995), our findings suggest that atomic speech, even without

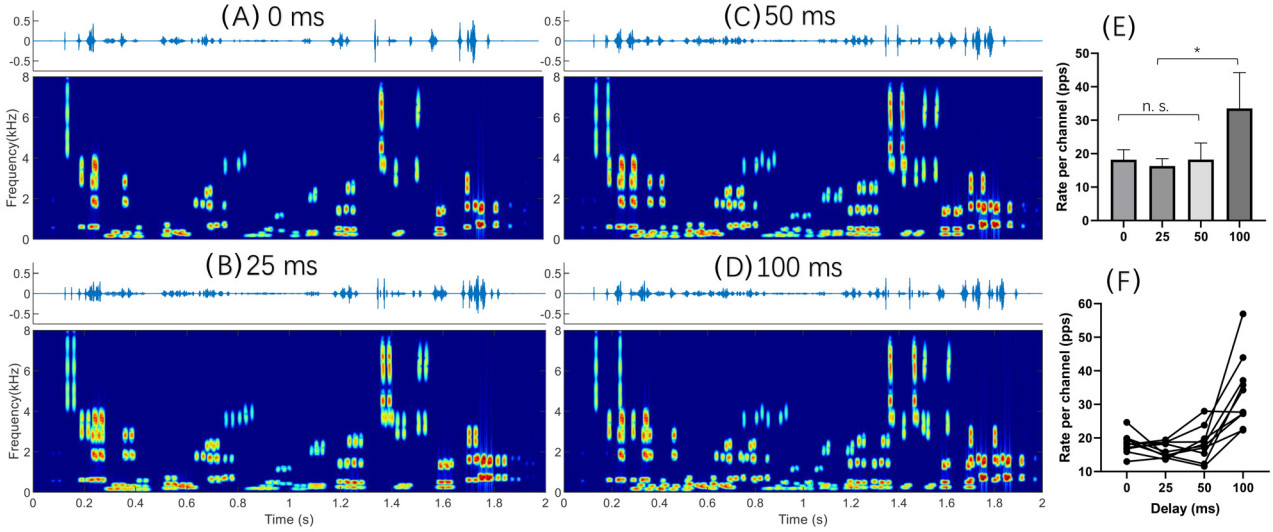

FIG. 6. SRT results (Rate per channel in pps. E: mean + SD with an asterisk indicating a significant difference; F: individual results) with atomic speech using four maxima with four echo delay conditions. Demonstrations of the four atomic speech signals are shown as follows: A) 0 ms, B) 25 ms, C) 50 ms, and D) 100 ms.

fine structure information and traditional acoustic cues, could still be adequate for speech comprehension.

In this experiment, we measured the SRTs in ten NH listeners with diotic atomic speech under three different spectral maxima conditions: one, four, and eight. The results showed that the mean SRT of 72.7 pps was the lowest with four maxima, compared to the other two maxima. Waveform analysis in Fig. 4(B) showed a predominantly pulsatile pattern in the atomic speech, contrasting with traditional continuous speech waveforms.

The most relevant findings in the literature come from a study on pointillistic speech (Kidd *et al.*, 2009). With 10-ms 16-tone pointillistic speech, approximately 70% intelligibility was achieved even when 87.5% of the short tones were randomly removed. In our experiments with atomic speech, 32 analysis bands were used. Under the 4-maxima condition, 87.5% (28 out of 32) of the spectral information was removed, coincidentally aligning with the percentage in the pointillistic speech. The mean 50% SRT per channel in our experiment 1 was ∼18.2 pps, corresponding to a frame duration of 55 ms. Although this frame duration is significantly longer than the 10 ms used in pointillistic speech, resulting in worse temporal resolution, atomic speech offers more focused Gabor tones in time compared to the short tones of pointillistic speech (see Fig. 1). Additionally, atomic speech demonstrates better spectral resolution (32 bands with selective 4-maxima extraction vs 16 bands with random tone removal in pointillistic speech). Because of the complex differences in implementation details between the two methods across the studies, the observed difference in mean intelligibility—from 70% for pointillistic speech to 50% for atomic speech—is reasonable and expected.

Furthermore, the atomic speech provides fresh insights into the concept of speech mode, akin to findings from studies on sine-wave speech (Remez *et al.*, 1981; Moore, 2013). The experiment task involved adaptive SRT measurement, where the initial trial in each run employed a high stimulation rate, resulting in relatively easy comprehension for participants. Subsequent trials featured decreasing rates, progressively elevating the task's difficulty. For inexperienced participants who had not taken part in the test, playing the atomic speech with a stimulation rate just above the threshold often resulted in descriptions like "bubbles" or "water-like" rather than recognizable speech. However, on being instructed to repeat the sentence, they transitioned into an irreversible speech mode and made efforts to recognize some of the content.

If we consider Gabor's suggestion that any sound can be broken down into a multitude of atoms (Gabor, 1947), the concept of atomic speech offers a distinct approach to exploring the impact of specific combinations of physical parameters on the perception of spectro-temporally sparse speech patterns. Building on the implementation of the four-maxima atomic speech in experiment 1, experiments 2 and 3 were conducted to investigate the influence of binaural integration and echo delay on the intelligibility of atomic speech.

B. Binaural integration of dichotic speech

For individuals with normal hearing, natural auditory input reaching both ears is continuous and highly correlated. Many works have demonstrated that humans can integrate speech acoustics cues presented dichotically to each ear (Cutting, 1976; Roberts and Summers, 2019; Roberts *et al.*, 2021; Sathe *et al.*, 2024). For instance, in Cutting (1976), listeners could report more than one voice but a fused message when only different fundamental frequencies were provided to each ear. In another line of research, Warren *et al.* (1995) measured the intelligibility of speech signals containing only two narrowband spectral slits. When presented individually, each band conveyed minimal information. However, when the two bands were presented simultaneously, intelligibility scores reached approximately 80%, regardless of whether the bands were delivered diotically (identical bands presented to both ears) or dichotically (one band assigned to each ear). Extending this paradigm, Spehar *et al.* (2008) compared monotic (two bands presented to one ear) and dichotic (bands split between ears) conditions. Their results showed that the dichotic condition yielded significantly lower mean percent content scores than the monotic condition—by 6.4% for young adults with normal hearing and 13.6% for older adults with normal hearing.

In experiment 2, we introduced a unique approach by randomly assigning atoms of the atomic speech to the left and right ears. Although this manipulation results in unnatural sounds, it offers a new opportunity to examine binaural integration abilities. The results indicated that, on average, participants faced significantly greater difficulty in the dichotic condition compared to the monaural condition. This is consistent with the NH results of Spehar *et al.* (2008). Interestingly, individual data in experiment 2 showed that some participants were able to effectively integrate information from both ears and perform as well as in the monaural condition. These findings underscore the variability in binaural integration among NH listeners. This suggests that factors beyond the scope of this study, such as cognitive abilities, musical training, and experience with other auditory tasks, may contribute to the differences in performance among individuals. Future research will aim to explore these potential correlations to better understand the mechanisms underlying individual differences in binaural integration.

In addition, Spehar *et al.* (2008) also compared the abovementioned monotic and dichotic conditions in an older adult hearing-impaired group. The mean scores were both around 33% with no significant differences. In another binaural way to use the interleaved band integration, studies involving hearing-impaired individuals using hearing aid (Kulkarni *et al.*, 2012) and CI users (Aronoff *et al.*, 2016), have shown that employing band-interleaved stimulation between the ears—where neighboring frequency bands are sent to opposite ears—can enhance speech understanding. This could be attributed to the limited spectral resolution and detrimental interchannel interaction experienced by the hearing-impaired listeners. Band-interleaved stimulation mitigates this interaction and provides more spectral

information to each ear. Consequently, in cases like these, the dichotic condition might prove more advantageous than the monaural or diotic conditions.

Our study, employing a unique stimulus pattern—atomic speech—provides further insight into binaural integration among NH listeners. Collectively, these results suggest that interleaved patterns across ears could yield both benefits (Aronoff *et al.*, 2016) and detriments (in current work and Spehar *et al.*, 2008) to speech understanding, and substantial individual variability may exist. These observations warrant further exploration in future studies.

C. Effects of single echo

In real-world settings, sound from a source can reach a listener's ears through various paths, involving direct and reflected routes. Sound reflections are a natural phenomenon that significantly contribute to the perception of direct sound. Reflected echoes are typically categorized into early reflections and late reflections, where the dividing point generally falls between 50 and 100 ms (Lochner and Burger, 1964; Warzybok *et al.*, 2013). Investigating the effects of real reflections or reverberation environments on speech intelligibility requires comprehensive studies, combining physical and psychoacoustic experiments, even when focusing on a single reflection (Rennies *et al.*, 2019). Additionally, understanding the underlying brain mechanisms is highly complex and demands further research (Gao *et al.*, 2024).

Experiment 3 investigated the impact of single echo on intelligibility of atomic speech. In this experiment, we introduced a delayed version of the atomic speech and combined it with the original direct atomic speech. The results indicated that a delay of 100 ms led to a significant increase in SRTs, whereas delays of 25 and 50 ms had no notable impact. This finding aligned with prior research, suggesting that late reflections can negatively affect the intelligibility of direct sound (Warzybok *et al.*, 2013; Li *et al.*, 2015). Early reflections can be integrated with the direct sound, whereas late reflections cannot and may even add masking effects. Binaural and temporal integration of reflections are crucial for related modeling work (Rennies *et al.*, 2019). Atomic speech is much sparser in time-frequency plane compared to natural speech. In our experiment, we tested the effect of adding a single echo of the atomic speech on the intelligibility. For non-echoic atomic speech, the mean SRT is 18.2 pps per channel, corresponding to a frame duration of approximately 55 ms. This suggests that around the mean SRT, echoes with delays within 50 ms often occur earlier than the Gabor tones in the next frame, whereas echoes with a 100 ms delay typically occur later than the Gabor tones in the subsequent frame, as illustrated in Fig. 6. These findings establish a connection between the SRT of non-echoic speech and echoic speech, and the results of atomic speech perception may provide insights into the perception of echoic intact speech.

These results showcased the atomic speech's potential to explore the influence of echoes on speech perception. Unlike typical continuous speech signals, atomic speech's sparsity in the spectro-temporal domain makes it conducive for quantifying and investigating interactions between direct and reflected atoms.

D. Speech sparsity models

A speech signal is sparse in the time-frequency plane, which facilitates listeners' abilities to glimpse target information from background noise. To investigate the contributions of energetic masking and informational masking, a series of studies have been performed using models such as the "glimpsing" model (Cooke, 2006) and ideal time-frequency segregation (Brungart *et al.*, 2006; Kidd *et al.*, 2019). The basic idea is to preserve or discard time-frequency points according to a predefined local signal-to-noise ratio criterion. Among their results, when the local signal-to-noise ratio criterion gradually increases above 0 dB, the spectrogram of the generated speech gradually became sparser and less masked by the interference. This paradigm could be used to quantify the effects of sparsity representation on speech perception.

Different from this series of work based on masking or computational auditory scene analysis, the atomic speech model treats any sound as a combination of a large number of Gabor atoms and produces a sparse signal by selecting a small portion of the atoms according to certain criteria, such as the time-frequency peaks in our specific implementations. Compared to the above studies using the criterion of local signal-to-noise ratio, the sparsity of atomic speech is not influenced by the selected masker and is purely determined by the characteristics of the target original speech itself.

Related to masker-independent sparsity models, there have also been many studies that quantitatively demonstrate their effectiveness such as sine-wave speech (Remez *et al.*, 1981), channel vocoded speech (Shannon *et al.*, 1995), mosaic speech (Ueda *et al.*, 2022), pointillistic speech (Kidd *et al.*, 2009). These studies distorted the speech in temporal, spectral, or spectro-temporal domains. The atomic speech model proposed in this study offers a new approach in this line to examine the sparsity of speech. All of these model studies quantitatively demonstrated speech redundancy by manipulating speech sparsity in specific ways. The speech recognition results were recorded as certain parameters changed, such as the number of formants, frequency channels, time window duration, etc.

E. Limitations and future work

To the best of our knowledge, this study is the first to describe the unique pattern of atomic speech and how well it can be understood. More research is necessary to explore how different combinations of settings affect the perception of atomic speech and its potential uses. The settings we chose are somewhat arbitrary and might need adjustments for specific situations. For instance, the rms normalization applied to atomic speech generation leads to higher energy for individual atoms as the overall rate decreases. This effect

could potentially hinder a fair comparison of the contributions of individual atoms to intelligibility in various atomic speech conditions. In addition, future studies should explicitly examine the relationship between the summation of monaural intelligibility performance in both ears and dichotic listening outcomes, as this aspect was not directly addressed in the current study, to provide stronger evidence for binaural integration and its underlying mechanisms. There could also be more creative and interesting patterns achievable using this framework. Further investigation into the mathematical and signal processing aspects of the atomic speech framework is also needed.

Different speech corpora are expected to yield varying mean atomic rate SRTs, influenced by factors such as rhythm, clarity, and articulation style. These variations would also affect results in binaural integration and echoic speech perception. For instance, slower original speech speeds (measured in syllables per second) may result in lower atomic rate SRTs as a result of the reduced original information rate. In the echoic speech test, if the reciprocal of the atomic rate SRT serves as a threshold for distinguishing early and late reflections, a lower SRT would correspond to a higher upper limit for early reflections. These hypotheses warrant further investigation.

In addition, the speech materials used in the experiments are in Mandarin Chinese, a tonal language. As we know, the tone information becomes increasingly critical for hearing-impaired listeners and under degraded conditions (Feng *et al.*, 2012; Chen *et al.*, 2014; Li *et al.*, 2019). During the experiment, the NH subjects did not report variations of lexical tone when they were able to understand the sentences. The representation in tone information in the atomic speech was not specifically tested in this study and warrants further investigation. A plausible hypothesis is that tone identification may be more challenging in isolated monosyllables than in natural phrases or sentences because of the influence of contextual cues. Additionally, exploring the effects of atomic speech processing in nontonal languages, such as English, would be a valuable avenue for future research.

V. CONCLUSION

This paper introduces the concept of atomic speech, a model that transforms original speech into a sparser signal, affecting the spectral and temporal domains. The study demonstrates that NH listeners can comprehend speech content with a total atom rate of under 80 pps based on our specific implementation, indicating high redundancy. The human auditory system may not need highly sampled speech signals for speech understanding. However, factors such as dichotic hearing and echo interaction may impede intelligibility or increase inter-subject variance.

ACKNOWLEDGMENTS

We express our gratitude to all volunteers for their participation and extend our appreciation to W. F. Liu for his valuable assistance in pilot experiments. This work received support from the Tertiary Education Scientific research project of Guangzhou Municipal Education Bureau (Grant No. 2024312312), Guangdong Basic and Applied Basic Research Foundation (Grant Nos. 2022A1515011361 and 2024A1515012585), and National Natural Science Foundation of China (Grant No. 12374448). During the preparation of this work, the authors used GPT-3.5 and GPT-4 language models to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

[1]See https://github.com/BetterCI/AtomicSound (Last viewed 26 February 2025).

Aronoff, J. M., Stelmach, J., Padilla, M., and Landsberger, D. M. (**2016**). "Interleaved processors improve cochlear implant patients' spectral resolution," Ear Hear. **37**, e85–e90.
Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.
Chen, F., Wong, L. L., and Hu, Y. (**2014**). "Effects of lexical tone contour on Mandarin sentence intelligibility," J. Speech. Lang. Hear. Res. **57**, 338–345.
Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.
Cooke, M., and Lecumberri, M. L. G. (**2020**). "Sculpting speech from noise, music, and other sources," J. Acoust. Soc. Am. **148**, EL20–EL26.
Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 597–606.
Cutting, J. E. (**1976**). "Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening," Psychol. Rev. **83**, 114–140.
Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.
Dudley, H. (**1939**). "Remaking speech," J. Acoust. Soc. Am. **11**, 169–177.
Feng, Y. M., Xu, L., Zhou, N., Yang, G., and Yin, S. K. (**2012**). "Sine-wave speech recognition in a tonal language," J. Acoust. Soc. Am. **131**, EL133–EL138.
Gabor, D. (**1947**). "Acoustical quanta and the theory of hearing," Nature **159**, 591–594.
Gao, J., Chen, H., Fang, M., and Ding, N. (**2024**). "Original speech and its echo are segregated and separately processed in the human brain," PLoS Biol. **22**, e3002498.
Guo, Z., Yu, G., Zhou, H., Wang, X., Lu, Y., and Meng, Q. (**2021**). "Utilizing true wireless stereo earbuds in automated pure-tone audiometry," Trends Hear. **25**, 23312165211057367.
Kidd, G., Jr., Mason, C. R., Best, V., Roverud, E., Swaminathan, J., Jennings, T., Clayton, K., and Colburn, H. S. (**2019**). "Determining the energetic and informational components of speech-on-speech masking in listeners with sensorineural hearing loss," J. Acoust. Soc. Am. **145**, 440–457.
Kidd, G., Jr., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., and Best, V. (**2016**). "Determining the energetic and informational components of speech-on-speech masking," J. Acoust. Soc. Am. **140**, 132–144.

Kidd, G., Jr., Streeter, T. M., Ihlefeld, A., Maddox, R. K., and Mason, C. R. (**2009**). "The intelligibility of pointillistic speech," J. Acoust. Soc. Am. **126**, EL196–EL 201.
Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.
Kong, F., Zhou, H., Mo, Y., Shi, M., Meng, Q., and Zheng, N. (**2023**). "Comparable encoding, comparable perceptual pattern: Acoustic and electric hearing," IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 2326–2337.
Kulkarni, P. N., Pandey, P. C., and Jangamashetti, D. S. (**2012**). "Binaural dichotic presentation to reduce the effects of spectral masking in moderate bilateral sensorineural hearing loss," Int. J. Audiol. **51**, 334–344.
Li, J. F., Xia, R. S., Fang, Q., Li, A. J., Pan, J. L., and Yan, Y. H. (**2015**). "Effect of the division between early and late reflections on intelligibility of ideal binary-masked speech," J. Acoust. Soc. Am. **137**, 2801–2810.
Li, N., Wang, S., Wang, X., and Xu, L. (**2019**). "Contributions of lexical tone to Mandarin sentence recognition in hearing-impaired listeners under noisy conditions," J. Acoust. Soc. Am. **146**, EL99–EL105.
Lochner, J., and Burger, J. (**1964**). "The influence of reflections on auditorium acoustics," J J. Sound Vib. **1**, 426–454.
Loizou, P. C. (**2006**). "Speech processing in vocoder-centric cochlear implants," Adv. Otorhinolaryngol. **64**, 109–143.
Lopez-Poveda, E. A. (**2014**). "Why do I hear but not understand? Stochastic undersampling as a model of degraded neural encoding of speech," Front. Neurosci. **8**, 348.
Lopez-Poveda, E. A., and Barrios, P. (**2013**). "Perception of stochastically undersampled sound waveforms: A model of auditory deafferentation," Front. Neurosci. **7**, 124.
Mandel, M. I., Grover, V., Zhao, M., Choi, J., and Shafer, V. L. (**2019**). "The bubble noise technique for speech perception research," Perspect. ASHA Spec. Interest Groups **4**, 1653–1666.
Mandel, M. I., Yoho, S. E., and Healy, E. W. (**2016**). "Measuring time-frequency importance functions of speech with bubble noise," J. Acoust. Soc. Am. **140**, 2542–2553.
Meng, Q. (**2020**). "Perception of atomic speech," J. Acoust. Soc. Am. **148**, 2722–2722.
Meng, Q., Zhou, H., Lu, T., and Zeng, F.-G. (**2023**). "Pulsatile Gaussian-enveloped tones (GET) for cochlear-implant simulation," Appl. Acoust. **208**, 109386.
Moore, B. C. J. (**2013**). *An Introduction to the Psychology of Hearing* (Brill, Leiden).
Nakajima, Y., Matsuda, M., Ueda, K., and Remijn, G. B. (**2018**). "Temporal resolution needed for auditory communication: Measurement with mosaic speech," Front. Hum. Neurosci. **12**, 149.
Patterson, R. D. (**2000**). "Auditory images: How complex sounds are represented in the auditory system," J. Acoust. Sci. Technol. **21**, 183–190.
Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (**1992**). "Complex sounds and auditory images," in *Auditory Physiology and Perception*, edited by Y. Cazals, K. Horner, and L. Demany (Elsevier, Oxford), pp. 429–446.
Popham, S., Boebinger, D., Ellis, D. P. W., Kawahara, H., and McDermott, J. H. (**2018**). "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," Nat. Commun. **9**, 2122.
Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–950.
Rennies, J., Warzybok, A., Brand, T., and Kollmeier, B. (**2019**). "Measurement and prediction of binaural-temporal integration of speech reflections," Trends Hear. **23**, 21–41.
Roads, C. (**2004**). *Microsound* (The MIT Press, Cambridge, MA).
Roberts, B., and Summers, R. J. (**2019**). "Dichotic integration of acoustic-phonetic information: Competition from extraneous formants increases the effect of second-formant attenuation on intelligibility," J. Acoust. Soc. Am. **145**, 1230–1240.
Roberts, B., Summers, R. J., and Bailey, P. J. (**2021**). "Mandatory dichotic integration of second-formant information: Contralateral sine bleats have predictable effects on consonant place judgments," J. Acoust. Soc. Am. **150**, 3693–3710.
Santi, Nakajima, Y., Ueda, K., and Remijn, G. B. (**2020**). "Intelligibility of English mosaic speech: Comparison between native and non-native speakers of English," Appl. Sci. **10**, 6920.
Sathe, N. C., Kain, A., and Reiss, L. A. J. (**2024**). "Fusion of dichotic consonants in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **155**, 68–77.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–303.
Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**, 87–90.
Spehar, B. P., Tye-Murray, N., and Sommers, M. S. (**2008**). "Intra- versus intermodal integration in young and older adults," J. Acoust. Soc. Am. **123**, 2858–2866.
Tang, Y. (**2022**). "The role of glimpses with different energy in speech-in-noise recognition," JASA Express Lett. **2**, 025201.
Ueda, K., Hashimoto, M., Takeichi, H., and Wakamiya, K. (**2024**). "Interrupted mosaic speech revisited: Gain and loss in intelligibility by stretching," J. Acoust. Soc. Am. **155**, 1767–1779.
Ueda, K., Takeichi, H., and Wakamiya, K. (**2022**). "Auditory grouping is necessary to understand interrupted mosaic speech stimuli," J. Acoust. Soc. Am. **152**, 970–980.
Vandali, A. E., Whitford, L. A., Plant, K. L., and Clark, G. M. (**2000**). "Speech perception as a function of electrical stimulation rate: Using the Nucleus 24 cochlear implant system," Ear Hear. **21**, 608–624.
Venezia, J. H., Hickok, G., and Richards, V. M. (**2016**). "Auditory 'bubbles': Efficient classification of the spectrotemporal modulations essential for speech intelligibility," J. Acoust. Soc. Am. **140**, 1072–1088.
Warren, R. M., Riener, K. R., Bashford, J. A., Jr., and Brubaker, B. S. (**1995**). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," Percept. Psychophys. **57**, 175–182.
Warzybok, A., Rennies, J., Brand, T., Doclo, S., and Kollmeier, B. (**2013**). "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," J. Acoust. Soc. Am. **133**, 269–282.
Wong, L. L., Soli, S. D., Liu, S., Han, N., and Huang, M. W. (**2007**). "Development of the Mandarin Hearing in Noise Test (MHINT)," Ear Hear. **28**, 70S–74S.
Xu, L., and Pfingst, B. E. (**2008**). "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," Hear. Res. **242**, 132–140.
Xu, L., Thompson, C. S., and Pfingst, B. E. (**2005**). "Relative contributions of spectral and temporal cues for phoneme recognition," J. Acoust. Soc. Am. **117**, 3255–3267.
Xu, L., Tsai, Y., and Pfingst, B. E. (**2002**). "Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses," J. Acoust. Soc. Am.. **112**, 247–258.
Xu, L., Xi, X., Patton, A., Wang, X., Qi, B., and Johnson, L. (**2021**). "A cross-language comparison of sentence recognition using American English and Mandarin Chinese HINT and AzBio sentences," Ear Hear. **42**, 405–413.
Xu, L., and Zheng, Y. (**2007**). "Spectral and temporal cues for phoneme recognition in noise," J. Acoust. Soc. Am. **122**, 1758–1764.