Gestalt Principles Emerge When Learning Universal Sound Source Separation

Han Li, Kean Chen, and Bernhard U. Seeber

***Abstract*—Sound source separation is an essential aspect in auditory scene analysis, which is still an urgent challenge for machine hearing. In this paper, a fully convolutional time-domain audio separation network (ConvTasNet) is trained for universal two-source separation, consisting of speech, environmental sounds, and music. Besides the separation performance of the network, the underlying separation mechanisms are our main concern. Through a series of classic auditory segregation experiments, we systematically explore the principles learned by the network for simultaneous and sequential organization. The results show that without prior knowledge of auditory scene analysis imparted on the network, it spontaneously learns the separation mechanisms from raw waveforms that are similar to those which have developed over many years in humans. The Gestalt principles for separation in the human auditory system are shown to be effective in our network: harmonicity, onset synchrony and common fate (coherent modulation in amplitude and frequency), proximity, continuity, similarity. The universal sound source separation network following Gestalt principles is not limited to specific sources and can be applied to various acoustic situations like human hearing, providing new directions for solving the problem of auditory scene analysis.**

***Index Terms*—Gestalt principles, separation mechanisms, universal source separation.**

I. Introduction

IN OUR daily lives, auditory scenes with multiple sound sources are ubiquitous. One of the most remarkable abilities of the human auditory system is to separate and track one source from complex scenes seemingly without effort. According to the seminal book of Bregman [1], auditory scene analysis (ASA) is based on two mechanisms, primitive and schema-driven grouping. The primitive grouping mechanism relies on intrinsic sound attributes (or cues) such as fundamental frequency, onset, loudness, etc., and is regarded as an innate, bottom-up process for simultaneous grouping and for binding components over time. Components in mixtures are separated and arranged into streams according to the Gestalt principles, including the principle of *proximity*, *similarity*, *continuation*, and *common fate*. On the other hand, the schema-driven mechanism represents top-down processing. Listeners exploit learned knowledge and attention to the further processing of complex auditory scenes.

Computational auditory scene analysis (CASA) models are technical source separation systems based on human auditory segregation principles. Many CASA systems model auditory scene analysis as a two-stage process: segmentation of time-frequency elements and grouping into auditory objects and streams [2]. Segmentation relies on the estimation of intrinsic, bottom-up sound attributes, such as pitch (e.g., Wang-Brown, 1999 [3]), amplitude modulation (Hu-Wang, 2004 [4]), or onsets (Hu-Wang, 2007 [5]). Then, according to continuity, synchrony, or other primitive grouping principles, these time-frequency segments are next grouped simultaneously across frequency and sequentially across time and frequency to produce auditory objects. Although these typical CASA models attempt to extract meaningful and biologically plausible cues, the accuracy of these cues, such as pitch or onset estimation in complex acoustic conditions, cannot be guaranteed, resulting in limited model performance.

With the development of neural networks, the approach to explicitly extract features has been gradually weakened. Also, the biological rational for modeling the process of auditory scene analysis is often no longer the main concern, but the improvement of separation performance for technical applications. Since these approaches do not mimic the auditory system's operation, they are not CASA models, but rather acoustic source separation or acoustic scene analysis approaches. Various supervised networks have been used in source separation and made great progress, especially for speech separation [6]–[8], such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs). Recently proposed end-to-end time-domain speech separation systems, such as Conv-TasNet [9] and DPRNN-TasNet [10], even surpassed the performance of ideal time-frequency masks.

Although performance has improved when using deep learning approaches, the underlying separation mechanisms of the network are more obscure. It is still unknown whether the separation is based on general primitive grouping principles, like in human hearing, or the pattern modeling of specific sound

Manuscript received June 25, 2021; revised December 25, 2021 and April 12, 2022; accepted May 2, 2022. Date of publication May 27, 2022; date of current version June 6, 2022. This work was supported by the TUM AIP through a 2-year Ph.D. scholarship for Han Li by the China Scholarship Council. The computer infrastructure was supported by the Bernstein Center for Computational Neuroscience, under Grant BMBF 01 GQ 1004B, and the Titan V card used was donated by the NVIDIA Corporation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jun Du. *(Corresponding author: Han Li.)*

Han Li is with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Audio Information Processing group, Department of Electrical and Computer Engineering, Technische Universität München, 80333 Munich, Germany (e-mail: lihan@mail.nwpu.edu.cn).

Kean Chen is with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: kachen@nwpu.edu.cn).

Bernhard U. Seeber is with the Audio Information Processing group, Department of Electrical and Computer Engineering, Technische Universität München, 80333 Munich, Germany (e-mail: seeber@tum.de).

Digital Object Identifier 10.1109/TASLP.2022.3178233

sources. If the network separates sounds based on general Gestalt principles like human hearing, this could be instructive for developing generalizing networks that do not depend on specific sound sources. In addition, it helps to explain the "black box" of deep networks, which is also an important issue that hinders the development of deep learning.

Few attempts have been made to interpret networks in biological terms. Francl and McDermott [11] demonstrated that trained networks can replicate key properties of mammalian spatial hearing, such as the sensitivity to monaural spectral cues and interaural time and level differences. Elhilali *et al.* [12] trained a hierarchical inference model to mimic the human auditory system for separation, demonstrating that some grouping principles are effective in this hierarchical inference model, such as harmonicity or frequency separation. However, due to the unsupervised learning, the model was not optimized for separation tasks, the accuracy for separation of actual complex sounds was not clear.

To our knowledge, except for our previous attempt to test some separation principles learned by the network [13], [14], there is no evidence that a supervised deep network for source separation can learn the Gestalt principles spontaneously like human hearing. In this study, we focus on the separation of two arbitrary sources in a monaural mixture consisting of a wide variety of sounds (speech, environmental sounds, music). A state-of-the-art end-to-end time-domain source separation framework, ConvTasNet, is adapted for separating arbitrary sources in Section III. In Section IV, the separation mechanisms are then explored through a series of classic auditory segregation experiments to test Gestalt principles. The experiments demonstrate, to our knowledge for the first time, that the Gestalt principles are intrinsically learned with supervised deep learning from unrelated natural sounds – a process and network that does not directly imitate the biological processing stages of the auditory system. The approach paves the way to establishing a universal separation network that can adapt to all scenes and achieve a segregation performance like in human hearing.

II. Auditory Grouping Principles

The 'Gestalt' concept originated in the 1920s [15] to explain visual object perception and was later extended to the auditory domain, c.f. the review by Bregman [1]. Combined with more psychological and physiological experiments, the Gestalt principles evolved into more specific principles for auditory scene analysis, such as harmonicity and onset synchrony. In the following, these specific Gestalt principles are introduced for simultaneous and sequential grouping.

For auditory perception, scene analysis can be divided into simultaneous and sequential organization, which depicts the processes for fusing and separating components across frequency and across time into one or multiple auditory objects, or sources. For simultaneous organization, there is a consensus that harmonicity and onset synchrony are the most important principles for segregating concurrent sounds in the auditory system [16]. In addition, common fate is also an important principle addressing the dynamic changes of concurrent sounds.

Harmonicity is a strong, common spectral regularity in natural sounds, especially in speech and music. It refers to the situation that frequencies of components are integer multiples of a common fundamental frequency (F0), which typically results from one single source. A wide range of psychoacoustic experiments on harmonicity for segregation and pitch perception has been conducted [17], showing that F0 differences aid concurrent sound segregation. Taking the identification of two sounds with different F0 as an example, experiments with double-vowels [18] and with orchestral instruments [19] show consistently that identification performance improves gradually as the F0 difference increases to two semitones and then asymptotes for further F0 difference increases [20].

If different frequency components change in the same way at the same time, they probably arise from the same source [1], [21] - the principle of "common fate" proposed by Gestalt psychologists [15]. Common fate in auditory scene analysis can be defined in terms of correlated changes in their amplitudes (amplitude modulation, AM) and their frequencies (frequency modulation, FM). AM refers to slow temporal fluctuations of the sound's intensity. Synchrony of the onset (common onset) is a special and critical example of AM, which has been shown to be one of the most powerful temporal principle for simultaneous component grouping [22]. When components share a common onset, it is likely that they have originated from the same source. On the contrary, components that start at sufficiently different times tend to be heard as separated sources. An onset asynchrony of about 30–50 ms is enough for affecting auditory grouping of pure tones [23] or the identification of double-vowels [24]. This principle is taken out from the common fate principle separately for detailed analysis.

The common fate principle here refers to coherent modulation in amplitude (AM) and frequency (FM). The role of AM for simultaneous grouping is common and useful; the modulation in speech caused by the opening and closing of the vocal cords contributes to the fusion of acoustic components [25]. When two tones are amplitude modulated by the same rate, they tended to be fused more strongly than when modulated with different rates. Moreover, components from the same source often share a common pattern of frequency modulation [26], which is thought to also contribute to fusion. Small fluctuations in frequency are common in speech and music instrument sounds, ranging from less than 1 percent to 10 percent of the carrier frequency, which is called "micromodulation" [1], [27], [28]. The micromodulation affects all frequency components that stem from one source, causing them to move in parallel and group into one coherent object.

The sequential organization is the process that assigns auditory time-frequency elements arriving sequentially over time to appropriate sources, which is often regarded as *auditory streaming* for the human auditory system [1]. For components in sequences, proximity, similarity, and continuity of their attributes are the most important principles for their sequential organization [29].

Proximity is the most intuitive and widely investigated Gestalt principle. It plays an essential role in auditory scene analysis, which refers to the proximity in frequency, time, loudness,

and other source attributes. In 1975, van Noorden [30] proposed the well-known temporal coherence boundary for auditory streaming based on frequency and temporal proximity. It shows that if the frequency and temporal distance between successive frequency components is large, they are more likely assigned to two streams by the auditory system.

The law of good continuation is a further Gestalt principle, which refers to the acoustic properties of components that are continuous or with a smooth transition, such as frequency or loudness. Any sequence that exhibits acoustic contiguity has probably come from one source. Abrupt changes in these attributes often mean the emergence of new sound sources.

The principle of similarity usually refers to a multidimensional sound attribute, timbre. Timbre is a complex auditory attribute relating to the spectro-temporal composition of stimuli that otherwise do not differ in pitch and loudness. It has been demonstrated that timbre dissimilarity can serve segregation [31]–[33].

After training the network, a series of experiments are conducted in Section IV to explore whether these specific Gestalt principles have been learned by the network.

III. Separation Network Model

A. Framework

With the development of deep learning, many source separation networks have made significant progress, especially for speech separation. However, few attempts have been made to separate arbitrary sources in monaural recordings [34], [35]. In this study, one end-to-end fully convolutional time-domain separation network (ConvTasNet) proposed by Luo *et al.* [9] is adapted to separate universal sound sources.

ConvTasNet follows the unified separation framework: encoder-separator-decoder. First, an encoder transforms the mixture waveform into intermediate representations by convolving with the framed mixture x with N encoding filters $\{h_n^{Enc}(t)\}_{n=0,\ldots,N-1}$ of length L:

$$\mathbf{X}(k,n) = \sum_{t=0}^{L-1} x(t+kH) h_n^{Enc}(t), \tag{1}$$

where $k \in \{0,\ldots,K-1\}$is the frame index of the waveform and H is the hop size. In this study, the encoder is freely learned through the training process by a 1-D convolutional layer. A rectified linear unit (RELU) layer is next applied to obtain nonnegative $\mathbf{X}^+(k,n)$ for the following separator.

The separator is used to estimate weighting functions (masks) for two sources through the time-dilated convolutional network (TDCN). It is chosen as three 1-D convolutional modules, and each module contains eight stacked 1-D convolutional blocks with different dilation factors. Other parameters are the same as the best non-causal model reported by Luo *et al.* [9]. The mask for the i-th source ($\mathbf{M}_i(k,n)$) is then multiplied with the mixture:

$$\mathbf{Y}_i(k,n) = \mathbf{X}^+(k,n) \odot \mathbf{M}_i(k,n), \tag{2}$$

where $\odot$ indicates point-wise multiplication. Reconstructed waveforms are calculated by transposed convolution with N decoding filters $\{h_n^{Dec}(t)\}_{n=0,\ldots,N-1}$ and overlap-add operation:

$$\hat{s}_i(t) = \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} \mathbf{Y}_i(k,n) h_n^{Dec}(t-kH). \tag{3}$$

B. Dataset

In this study, we attempt to train a universal network to adapt to various acoustic scenes. Therefore, we used a universal dataset to train ConvTasNet, including environmental sounds (e.g., vehicle noise, bells, animal calls, etc.) from the BBC sound effects dataset [36], speech from the LibriSpeech database [37], and music without vocals from the musan database [38]. For data pre-processing, files only with background noise or with multiple overlapping sounds were excluded. All segments were downsampled to 16 kHz and cut to 3 s length. Environmental sounds, speech, and music had the same proportion in the dataset.

To create mixtures, two source clips were chosen randomly and mixed with random signal-to-noise ratios (SNRs) between -5 dB and $+5$ dB. To avoid confusion, mixing from the same sound source was not allowed, such as the same speaker, the same music track, or the same class (e.g., cars) in the environmental sound. Overall, the dataset included 180000 clips (150 hours), of which 70% were randomly selected for training (105 hours), 20% for cross-validation (30 hours), and 10% for testing (15 hours).

C. Training and Evaluation Setup

The scale-invariant source-to-distortion ratio (SI-SDR) [39] is used as an objective training target and measure of separation accuracy. It directly calculates the fidelity in the time domain by comparing the given true source s and the estimated source $\hat{s}$, which can be expressed as

$$\mathrm{SI-SDR}(s,\hat{s}) = 10\log_{10}\frac{\|\alpha s\|^2}{\|\alpha s - \hat{s}\|^2}, \tag{4}$$

where $\alpha = <s,\hat{s}>/\|s\|^2$, and $<>$ indicates the inner product. SI-SDR improvement (SI-SDRi) is the difference between output SI-SDR and input SI-SDR, where the network output signal and input mixture signal are regarded as $\hat{s}$ to calculate output SI-SDR and input SI-SDR through (4), respectively.

Permutation invariant training (PIT) [40] is adopted to address the source label permutation problem, which aligns the network output and the given true source during training. All possible assignments between estimated and clean sources ($\hat{s}_1 \sim s_1$, $\hat{s}_1 \sim s_2$, $\hat{s}_2 \sim s_1$, $\hat{s}_2 \sim s_2$) are listed. Then the SI-SDR is calculated for each assignment to get the pairwise scores. The maximum score for different assignments is chosen as the training objective.

All experiments are implemented with the Asteroid toolkit [41] and are trained through the Adam optimizer for 100 epochs.

TABLE I
SI-SDRi (dB) for Mixtures From Different Source Types and Average in the Test Dataset

Method	window size	speech + speech	environmental sounds + environmental sounds	music + music	speech + environmental sounds	speech + music	environmental sounds + music	AVG
ConvTasNet	**2 ms**	**13.45**	**9.84**	**5.81**	**13.41**	**14.28**	**10.34**	**11.70**
	4 ms	12.40	9.50	5.28	12.77	13.04	9.60	10.89
	8 ms	10.99	9.34	4.93	11.70	11.56	8.72	9.91
	16 ms	8.79	9.17	4.62	11.07	10.73	8.73	9.28
	32 ms	7.54	7.37	4.44	9.51	10.12	7.77	8.24
IRM	2 ms	9.31	10.02	6.01	9.82	7.85	8.54	8.61
	4 ms	9.59	11.09	6.73	10.52	8.46	9.51	9.34
	8 ms	9.81	12.17	7.56	11.19	9.15	10.53	10.10
	16 ms	10.99	13.18	8.47	12.28	10.41	11.59	11.20
	32 ms	**12.97**	**14.00**	**9.70**	**13.58**	**12.17**	**12.71**	**12.59**

Where Best Results of ConvTasNet and IRM are Shown in Bold.

D. Results

The separation performance for the test dataset is shown in Table I. As mentioned before, our model is used for a universal dataset, which includes speech, environmental sounds, and music. In the following, not only the average results but also the results for different source types are presented.

The encoder in ConvTasNet is framewise, where the kernel size (also called window size) in the 1-D convolutional layer controls the frame rate, which determines the context viewed by the network and is an important parameter for separation performance. The results of the ideal ratio mask (IRM) are calculated for comparison. IRM indicates the ratio of the target source energy to mixture energy in spectro-temporal units. It is one commonly used indicator of the dataset difficulty. Because the IRM is based on the spectrogram, the STFT is also calculated with different window sizes for comparison.

For ConvTasNet, the best average SI-SDRi is 11.70 dB and obtained when the window size is 2 ms. Compared with the IRM, the network shows promising results. In general, the performance is different for different source types. The separation of speech outperforms others, where speech and music separation is 14.28 dB, followed by speech and speech separation (13.45 dB), then speech and environmental sounds separation (13.41 dB). They are comparable or even surpass corresponding best IRMs. This may be due to the unique harmonic structure of speech which is more easily learned by the network. In contrast to speech separation, the performance for other mixture types is not satisfactory, where separation of music from music is the worst. A piece of music in the dataset is not played by only one instrument. It includes many tracks with various instruments, such as piano, drums, bass *et al.* When different music pieces are mixed, especially for similar music genres, it will be hard to separate all different instruments into the respective source signal mixtures. The IRM results of music and music separation are also the worst among other mixtures, indicating greater difficulty. Different types of sound sources have their own unique characteristics and may affect the learning of principles, which could be interesting for future study.

The performance of ConvTasNet decreases as window size increases, which is opposite to the trend of the IRM. The introduction of dilated convolution with increasing dilation factors in the network ensures that neurons in the highest layer of ConvTasNet can be affected by a long enough context [42]. A smaller window size allows the network to have a higher temporal resolution for each frame, which improves separation performance. The effect of window size on performance varies with the type of source. The separation of speech and other sources is sensitive to the window size, while environmental sounds and music are less so.

In addition to the average values in Table I, scatter plots are provided to see the distribution of model separation results. Fig. 1 shows scatter plots of input SI-SDR and SI-SDRi (dB) of the results of ConvTasNet with 2 ms window size for mixtures from different source types in the test dataset. The color scale is the density estimated by Gaussian kernel density estimation. The results generally show a downward trend, indicating that for lower input SI-SDR it is easier to obtain a larger SI-SDRi, which was previously observed [43]. In addition, the distribution of speech and other sources separation is more compact with a higher cluster center and fewer failure cases, while it is less concentrated for separations of environment sounds and music.

Taking the separation of speech and dog barking as an example, Fig. 2 shows the spectrogram of the mixture, both sources, and the two separated outputs of the network. The SI-SDRi for both sources are 12.11 dB and 9.81 dB, respectively. It can be seen from the figure that except for some instantaneous components, these two sources are well separated and reconstructed.

IV. Separation Mechanisms for Simultaneous and Sequential Organization

The trained ConvTasNet model achieved good performance for universal source separation. But what are the underlying separation mechanisms? Is separation based on modeled patterns of specific sound sources or on generalized primitive grouping principles?

In perception, auditory scene analysis can be divided into simultaneous and sequential organization. Simultaneous organization forms an object from concurrent components across frequency, while sequential organization links components across time. For the auditory system, the main grouping principles

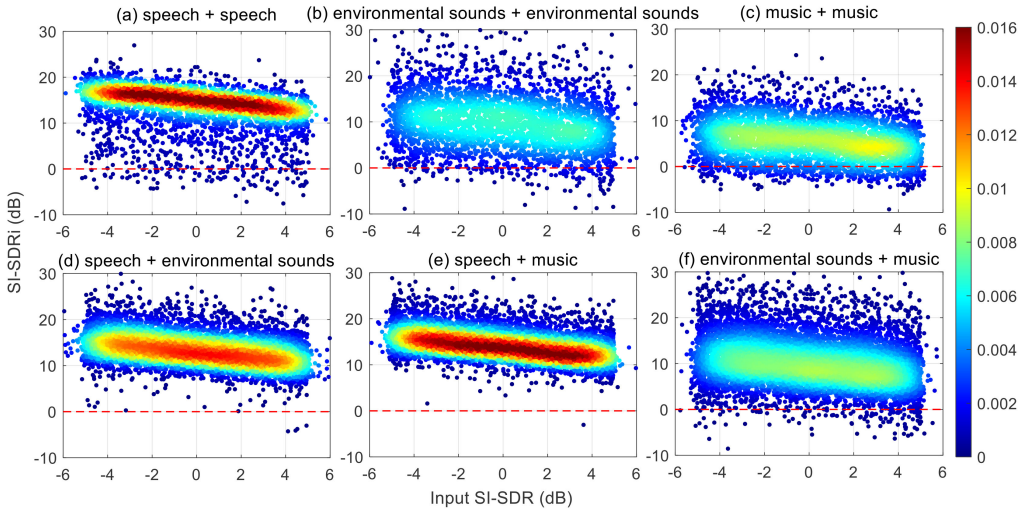


Fig. 1. Scatter plots of input SI-SDR and SI-SDRi (dB) for different source types in the test dataset, where (a)–(f) show different combinations of speech, environmental sounds, and music, respectively. Warmer colors indicate higher density, which is estimated by Gaussian kernel density estimation.

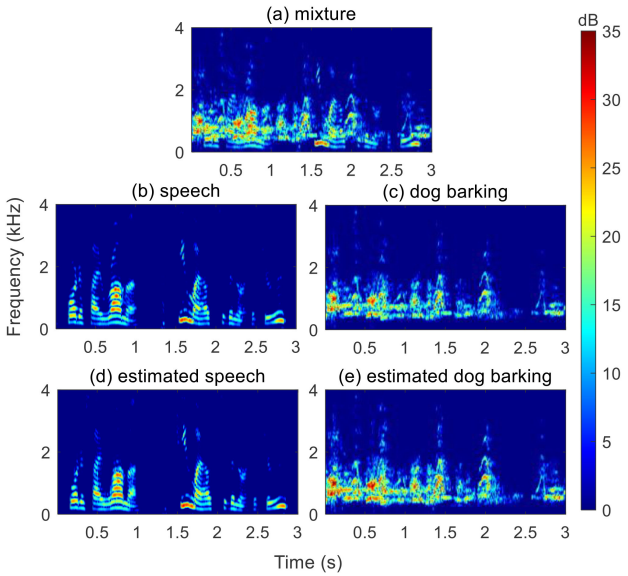


Fig. 2. The spectrogram of (a) mixture, (b)–(c) given true sources, and (d)–(e) estimated sources of the network, where two sources are speech and dog barking. Because the energy above 4 kHz is small, for better readability the maximum value of the y-axis is set to 4 kHz.

for simultaneous organization are harmonicity, onset synchrony, and common fate (AM and FM).

The proximity in frequency and time, the similarity in timbre, and continuity are predominant in the sequential organization. In this paper, we chose a set of classic segregation experiments probing simultaneous and sequential organization to test the model's functioning.

A. Methods and Stimuli

The Gestalt principles are often investigated one by one through artificial stimuli composed of discrete frequency components in most research on ASA [1]. We follow this approach with a series of experiments from classic ASA literature using two sound stimuli summed in the one input channel of the trained network.

Because the network is trained with a universal sound source dataset and has been verified that the network is capable of separating complex natural sources. Artificial stimuli that the network has not seen before are used to test the network's underlying mechanisms. These artificial stimuli completely differ in their kind and their spectral and temporal composition from the training dataset. Only when networks generalize segregation principles, the separation performance of untrained artificial stimuli may follow that of human auditory scene analysis.

All experiments are conducted on the best model (ConvTasNet with 2 ms window size) trained by the universal dataset from Section III-B without any other adjustment. The network separation result (SI-SDRi) is used as a performance indicator to analyze the principle's effect.

Two types of stimuli are used in the simultaneous and sequential organization experiments, respectively. The stimuli used for probing simultaneous organization are two harmonic complexes with different fundamental frequencies (F0) with or without common onset, as shown in Fig. 3(a). The duration of the stimulus is 3 s, including 200 ms raised-cosine onset and offset ramps to reduce transient effects. Each component in the mixture has equal amplitude, which means that the SNR of the two sources is equal to 0. There are three harmonics in one source, and $F0_1$ is fixed at 110 Hz. Here, four simultaneous organization experiments are conducted to test harmonicity, onset synchrony, and common fate (AM and FM), respectively.

For experiment 1, the stimuli are two groups of harmonics with different F0 and with common onset. F0 differences (ΔF0s) are varied from 0 to 12 semitones in steps of 0.1 semitones. The semitone scale is adopted in this paper because it is commonly used in psychoacoustic experiments [44], and the perceived pitch of complex tones is generally proportional to the logarithm of the frequency. Each semitone is one-twelfth of an octave, and an

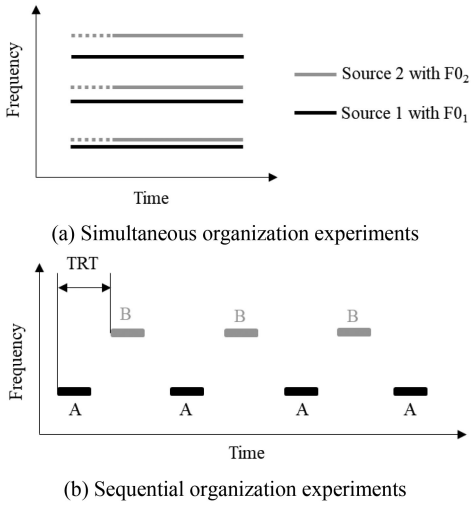


Fig. 3. Schematic spectrogram of the stimuli used in the experiments. (a) Simultaneous organization experiments, where black and gray indicate two concurrent sound sources. They are two harmonic complexes with different fundamental frequencies. (b) Sequential organization experiments, where A and B represent alternating sources appearing in sequence. They may be pure tones or harmonic complexes, differing in parameters like fundamental frequency, timbre, or any other property.

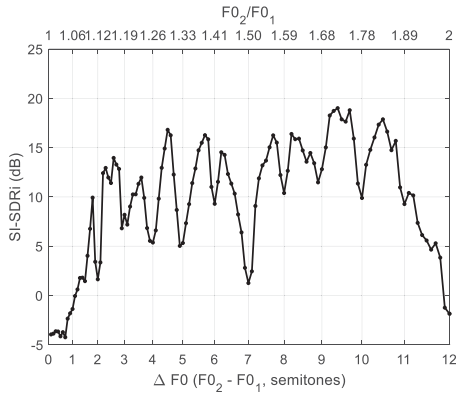


Fig. 4. Harmonicity experiment: SI-SDRi (dB) as a function of F0 differences between two complex tones, where ΔF0 (F0$_2$–F0$_1$, semitones) is marked on the bottom x-axis, and the ratio of F0$_2$ to F0$_1$ is marked on the upper x-axis.

octave is an interval between one tone and another with double its frequency. When f_2 is one semitone higher than f_1 , $f_2 = f_1 \times 2^{1/12}$.

Then we introduce onset asynchrony into the harmonicity experiment to study the contribution of onset asynchrony to segregation. For experiment 2, Δonset is varied from 0 s to 1.5 s (half of the duration of source 1) in steps of 0.01 s, where Δonset indicates the delay of source 2 from source 1.

Finally, we conducted two experiments to test whether the introduction of AM (experiment 3) and FM (experiment 4) with different modulation depths and rates contribute to separation. The parameters (ΔF0 and Δonset) that have been tested in the above experiments are fixed, and a case with partial segregation based on harmonicity is chosen here (ΔF0 = 1.5 semitones, and Δonset = 0 s). Source 1 is unmodulated as before and all three components in source 2 are sinusoidally modulated in amplitude or in frequency.

In the AM experiment, the modulation depth is varied from 0 to 100% in steps of 2%, and the modulation rate is changed from 0 to 5 Hz in steps of 0.1 Hz. In the FM experiment, because micromodulation is the common pattern of frequency modulation in natural sounds and can be perceived by the auditory system, the modulation depth here is varied from 0 to 10% in steps of 0.2%. The modulation rate is varied from 0 to 5 Hz in steps of 0.1 Hz.

For sequential organization experiments, a classic stimulus paradigm proposed by van Noorden [30] for auditory streaming is used, including two alternating components A and B with different frequency and tone repetition time (TRT(ms), the onset to onset time for two adjacent tones), as shown in Fig. 3(b). A and B are 40 ms in duration, including 5 ms raised-cosine onset and offset ramps to reduce transient effects. Each sequence consists of 10 A-B components in total. A and B can be pure tones or harmonic complex tones. In the following, three experiments are conducted to investigate whether proximity, continuity, and similarity principles emerge through model training.

For the proximity experiment, the proximity in frequency and time is investigated. A and B are pure tones. B-tones are fixed at 1 kHz, and A-tones are varied from 0 to 15 semitones below B tones in steps of 0.2 semitones. TRT varies from 50 to 200 ms in steps of 2 ms. These parameters are replicated from the classic psychoacoustic experiment of van Noorden [30] to compare the model's behavior with human performance.

For the continuity experiment, we introduce smooth frequency transitions between successive tones on the basis of the proximity experiment, which changes "discrete" tones A and B to be "connected". The samples in frequency transitions are generated through a logarithmic swept-frequency cosine signal [45], where the start and end frequency are frequency of tone A and B, respectively.

For similarity in the timbre experiment, due to the multidimensional nature of timbre, it is difficult to quantify the similarity in timbre space since the relationship of components in amplitude, temporal, and spectral spacing contribute. Here, the timbre differences stem from using different sets of three adjacent harmonics. Two alternating sources A and B with three harmonics to the same fundamental frequency (F0 = 110 Hz) and with equal amplitude are presented which provide the same pitch. A set of 10 timbres (T1–T10) with different harmonic numbers are created, where T1 uses harmonics 1, 2, and 3, T2 uses harmonics 2, 3, and 4, and T10 uses harmonics 10, 11, and 12. The harmonics have a duration of 250 ms with 20 ms raised-cosine onsets and offsets ramps and the TRT is 350 ms.

B. Results of Simultaneous Organization Experiments

1) Harmonicity: For the first experiment probing segregation of simultaneous harmonic complex tones by differences in fundamental frequency, results are shown in Fig. 4 as SI-SDRi (dB) as a function of F0 differences, where ΔF0 (F0$_2$–F0$_1$, semitones) is marked on the bottom x-axis, and the ratio of F0$_2$ to F0$_1$ is marked on the upper x-axis. When ΔF0 < 2 semitones, the network separation performance increases as ΔF0 increases. When

ΔF0 is at 2–10 semitones, as ΔF0 increases, the separation performance fluctuates with one semitone period. When ΔF0 is at 10–12 semitones, the first harmonic of $F0_2$ is close enough to the second harmonic of $F0_1$, and it tends to be perceived as a unitary source again.

The symmetric increase and decrease at 0–2 and 10–12 semitones result in the same conclusion that ΔF0 contributes to the separation of the sound source, which is consistent with the auditory system [46]. Within a certain range, the increase of ΔF0 significantly promotes segregation, such as the 0–2 semitones range for double-vowel recognition [18], and fewer further improvements beyond this range can be obtained.

When ΔF0 exceeds that range, that is, for 2–10 semitones difference, the difficulties to separate two sources when ΔF0 is an integer multiple of semitones are also evidence for the role of harmonicity for component grouping. In these cases, all the components from sources 1 and 2 are integer multiples of semitones, they are more harmonic, resulting in grouping into one source. It is evidence for supporting that the semitone is the smallest interval commonly used in Western tonal music [47]. If tones are not in the semitone scale, they tend to be segregated and perceived as dissonant and unpleasant. Further, the result is consistent with the theory of musical consonance [48]–[50], which is an auditory perceptual phenomenon that simple frequency ratios between two tones give much higher consonance than other ratios. The most consonant intervals are with ratios 1:2 ($F0_1 = 110$ Hz, $F0_2 = 220$ Hz, $\Delta F0 = 12$ semitones), 2:3 ($F0_1 = 110$ Hz, $F0_2 = 165$ Hz, $\Delta F0 = 7$ semitones), 3:4 ($F0_1 = 110$ Hz, $F0_2 = 146.8$ Hz, $\Delta F0 = 5$ semitones), 4:5 ($F0_1 = 110$ Hz, $F0_2 = 138.6$ Hz, $\Delta F0 = 4$ semitones), which correspond to the local minima of SI-SDRi in Fig. 4 that are more difficult to separate.

The input and outputs of the network for two stimuli with different a) $\Delta F0 = 7$ semitones and b) 7.7 semitones are shown in Fig. 5 to explain how the network exploits the harmonic constraints to assign these six frequency components to one or two sources. The worst separation performance in 2–10 semitones is reached for $\Delta F0 = 7$ semitones, a fifth in music and a highly consonant tone combination widely used in music ("power chord"). As shown in Fig. 5(a), the 3rd harmonic of source 1 is identical to the 2nd harmonic of source 2, where $F0_1 = 110$ Hz and $F0_2 = 165$ Hz. All components with the most energy in the mixture are assigned to one estimate. The energy in the other estimate is small, but some components emerge that do not exist in the stimuli, a series of harmonics of 55 Hz. It indicates that components are estimated by the network based on harmonic constraints. When the component frequencies of two sources share harmonic relationships, they are likely be regarded as the same sound source and even evoke virtual fundamental frequencies [48].

Another example that harmonicity is exploited to separate two sources is $\Delta F0 = 7.7$ semitones, as shown in Fig. 5(b). When ΔF0 is not an integer multiple of semitones, there is no specific harmonic relationship between F0 of the two sound sources. Then, according to the harmonicity between the three components of each sound source, all components can be correctly assigned to two sources.

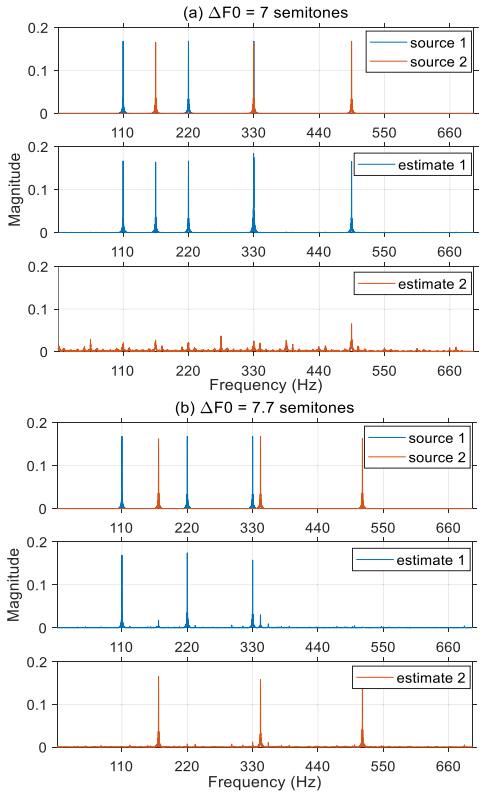


Fig. 5. The input and outputs of the network for two stimuli with different ΔF0, 7 semitones (a) and 7.7 semitones (b). The top panels show the spectrum of the input mixture, and the middle and bottom panels show the two sources' estimates.

In general, for concurrent components, harmonicity is learned by the network and used effectively for separation. The harmonicity principle exploited by the network is consistent with the auditory system in the following aspects. First, within the range of 2 semitones, ΔF0 contributes to source separation. Second, components are always assigned due to harmonic constraints. Based on whether the harmonicity is within each source or across two sources, the harmonicity has two effects on separation, beneficial or hindering. When the fundamental frequencies of the two sound sources have no harmonic relationship, that is, only components belonging to the same source are harmonic, they will be correctly separated into two sound sources due to the harmonicity within each source. When ΔF0 is an integer multiple of semitones, all components from two sources share a harmonic relationship. At this time, the harmonicity helps to combine all components into a single source and hinders the separation of the two sound sources.

2) Onset Synchrony: Results of SI-SDRi (dB) as a function of Δonset for $\Delta F0 = 1.6$, 7.0, 7.7 semitones are shown in Fig. 6. For $\Delta F0 = 1.6$ semitones, a condition which shows some segregation already for simultaneous onsets, SI-SDRi increases with the increase of Δonset. Onset asynchrony is segregation-promoting, consistent with the auditory system. Specifically, SI-SDRi increases rapidly when a slight delay ($<$100 ms) breaks the synchronization of both sources, and then increases gradually until it approaches the asymptote. Asynchrony of more than

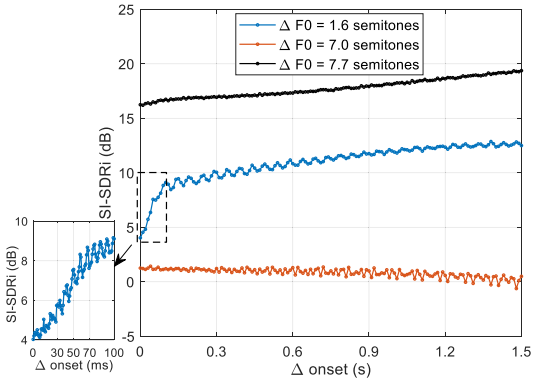


Fig. 6. Onset synchrony experiment: SI-SDRi (dB) as a function of Δonset (s) between two complex tones for ΔF0 = 1.6, 7.0 and 7.7 semitones, where the first 100 ms for ΔF0 = 1.6 semitones are zoomed in.

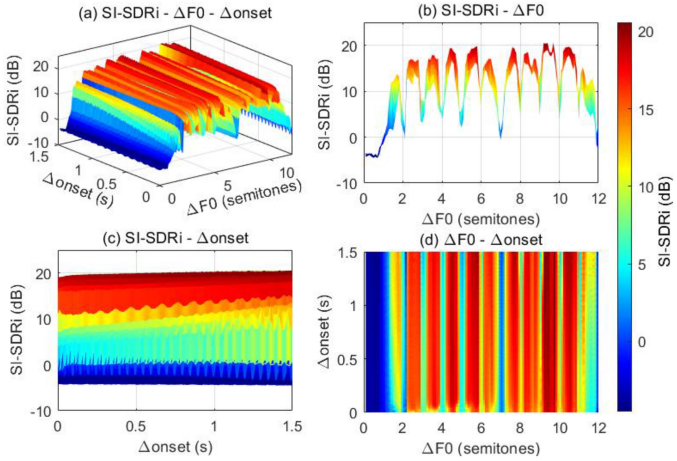


Fig. 7. Onset synchrony experiment: SI-SDRi (dB) as a function of ΔF0 (semitones) and Δonset (s): (a) 3D view, (b) projection on the SI-SDRi–ΔF0 plane, (c) projection on the SI-SDRi–Δonset plane, and (d) projection on the ΔF0–Δonset plane. Warmer colors indicate a better separation.

30 ms has been shown to be helpful for auditory separation [51], and similarly for ConvTasNet, the impact of 10–20 ms Δonset is also relatively small while for Δonset of 30–50 ms a rapid increase in segregation occurs. For ΔF0 = 7.7 semitones, a condition that is already well segregated with simultaneous onsets, the contribution of Δonset is limited. However, when ΔF0 is in integer multiples of semitones (ΔF0 = 7.0 semitones), onset asynchrony-based segregation is not sufficiently powerful to overcome grouping due to harmonicity with ConvTasNet. Regardless of Δonset, the network tends to fuse all components into one source. In this case, harmonicity has a greater weight than onset asynchrony for the network.

Having seen that both harmonicity and onset synchrony contribute to segregation by the network, we are interested in their detailed interaction. Fig. 7 shows separation performance as a function of both parameters. As visible in Fig. 7(b), the introduction of Δonset does not change the trend between SI-SDRi and ΔF0. The separation performance is dominated by ΔF0, whereas onset asynchrony contributes only in cases with partial segregation based on F0, as one would expect. The

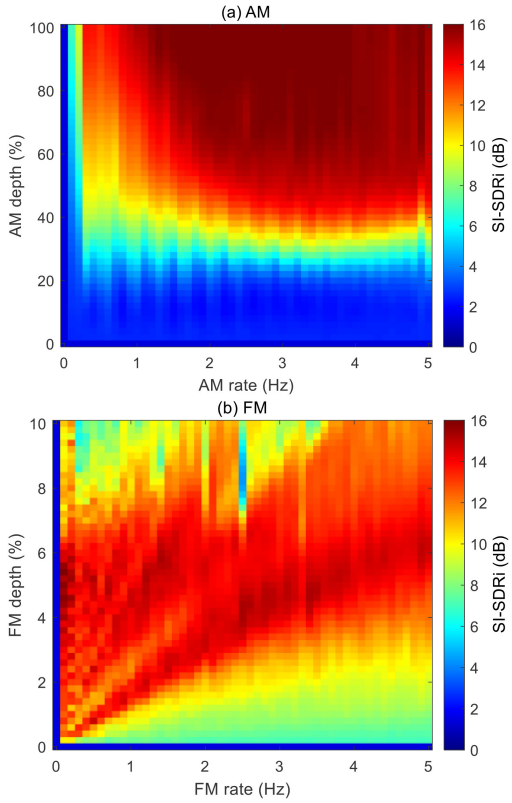


Fig. 8. Common fate experiment: SI-SDRi (dB) as a function of modulation rate and depth in amplitude modulation (a) or in frequency modulation (b).

contribution of Δonset is different for different F0, and three typical cases have been analyzed in Fig. 6. The projection on the ΔF0–Δonset plane in Fig. 7(d) indicates that harmonicity and onset asynchrony contribute almost independently.

In summary, for the network, harmonicity is the overall dominant principle in inducing simultaneous segregation, but onset asynchrony also facilitates segregation, especially for conditions with partial separation by harmonicity. These two principles are processed almost independently.

3) Common Fate: Results of experiment 3 (AM) and experiment 4 (FM) are given in Fig. 8(a) and (b), respectively. For the AM experiment, the introduction of low-rate AM into one of the two sources helps to group these three harmonic components that share the same modulation, resulting in increased separation. The thresholds of AM depth and rate that effectively promote separation are about 30% and 0.3 Hz for this experiment. Beyond this threshold, separation performance increases with the increase of modulation depth and rate. Segregation appears to peak at around 3–4 Hz, the syllable rate of speech, which also agrees with the maximum perceived fluctuation strength [52].

For the FM experiment, the introduction of micromodulation bring obvious benefits for separation. It provides new support for grouping three harmonics that share the same pattern of fluctuation in frequency in addition to the harmonicity principle. Segregation peaks at a modulation depth of 3%–5%. For the auditory system, it is the effective range of FM in voiced portions of speech [27].

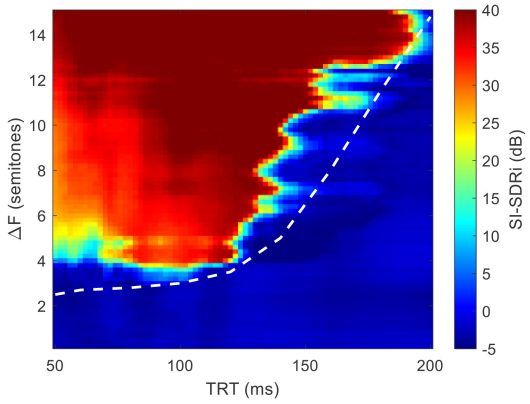


Fig. 9. Proximity experiment: SI-SDRi (dB) for tone sequences separation as a function of ΔF (semitones) and TRT (ms) in a 2D color map. The temporal coherence boundary obtained by van Noorden's psychoacoustic experiments [30] is indicated by the white dashed line.

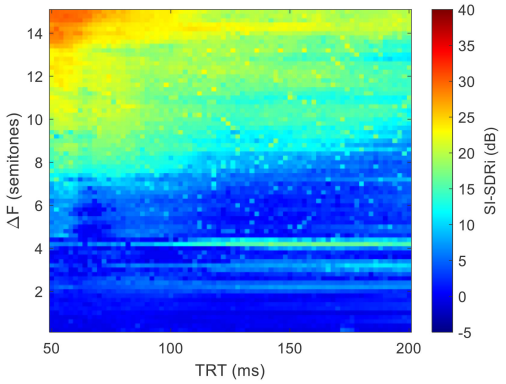


Fig. 10. Continuity experiment: SI-SDRi (dB) for tone sequences separation as in Fig. 9, but with a smooth frequency transition, where warmer colors indicate a better separation between A and B tones.

In general, a difference in amplitude and frequency modulation assists source separation. These phenomena are consistent with the auditory system, and it is plausible to believe that common fate is learned effectively by the network.

C. Results of Sequential Organization Experiments

1) Proximity: Results of the experiment with alternating tone sequences are given as SI-SDRi (dB) as a function of ΔF and TRT in Fig. 9. When ΔF is large and TRT is short (top left corner of Fig. 9), tone sequences A and B are more likely to be separated. On the contrary, the proximity in frequency and time will hinder model separation. It is consistent with the temporal coherence boundary presented by van Noorden [30]: when the tone interval is higher than the temporal coherence boundary (the white dashed line in Fig. 9), listeners tend to perceive two sound sources. It appears that the proximity principle in frequency and time is learned automatically by the network with a similar parametric outcome as in humans.

2) Continuity: Results of continuity experiment are shown in Fig. 10, again as SI-SDRi (dB) as a function of ΔF and TRT. The contribution of continuity, introduced in the experiment by linking A and B tones by logarithmic sweeps, is understood by comparison against the proximity experiment with discrete

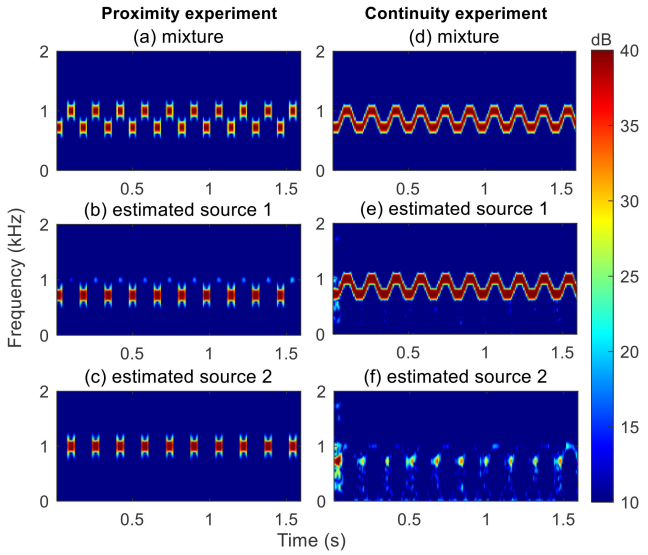


Fig. 11. Spectrogram of the mixture (top row) and two sources estimated by the network (rows 2 and 3) for ΔF = 5.6 semitones and TRT = 0.08 s. The left column shows in panels (a) (b) (c) results from the proximity experiment, in which A-B tones alternate, while the right column, panels (d) (e) (f), shows results from the continuity experiment in which A-B tones were connected with logarithmic sweeps.

frequency jumps. Compared with Fig. 9, the results in Fig. 10 share the same general trend that when ΔF is large and TRT is short (top left corner of Fig. 10), mixtures are more likely to be separated. More importantly, regardless of the interval in the time-frequency domain, the separation performance significantly dropped after the introduction of a smooth transition. The good continuation hinders sequential segregation effectively.

The spectrogram of the mixture and the two sources estimated by the network for ΔF = 5.6 semitones and TRT = 0.08 s are shown in Fig. 11 for direct comparison without (left column, proximity experiment) and with (right column, continuity experiment) continuation. Without continuation, the model entirely separates the mixture into two sources of high and low frequencies. On the contrary, mixtures are grouped into one source by the network when consecutive tones are connected with the frequency transitions.

3) Similarity: Results for sequential segregation based on timbre are shown in Fig. 12. When source A and B are more similar in timbre, that is, harmonics of source A and B with the same F0 are in a closer frequency region, they are more likely to integrate into one stream. On the contrary, when the differences in spectral spacing are larger (upper-left and lower-right corner), they are separated into two sources by the network despite sharing the same fundamental frequency and intensity. It indicates that the timbre differences, here from differences in the spectral centroid, can be learned by the model and are used effectively for separation as an additional cue.

V. Competition and Cooperation of Simultaneous and Sequential Organization

If a group of sound components can be regarded as arising from the same physical source, they should have a simultaneous or sequential relationship. As mentioned in the above sections,

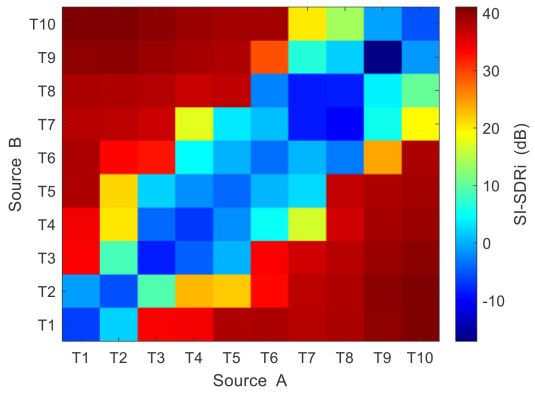


Fig. 12. Similarity experiment: SI-SDRi (dB) for separation of alternating complex tones A and B with different timbre, but identical F0.

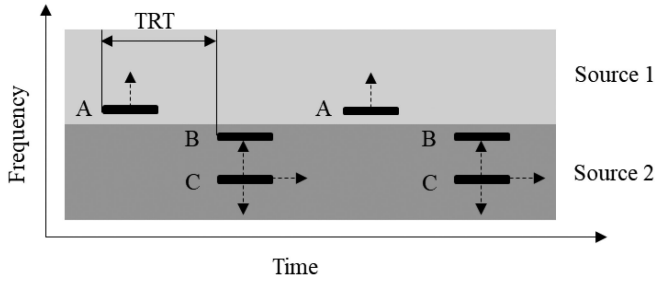


Fig. 13. Schematic spectrogram of the stimuli used in this experiment, replotted after Bregman and Pinker, Fig. 1 [23]. There are alternating tones A, B and C, where simultaneous integration tends to group B and C tones and sequential integration tends to group A and B tones. Here, this separation pattern (sequences A is regarded as source 1 and complex tones B and C is source 2) is used to calculate the SI-SDRi of the network. The dashed arrows indicate the change direction of tones.

there are various factors that promote the appropriate separation of mixtures. Harmonicity, onset synchrony, and common fate (AM and FM) are predominant for the simultaneous organization. The sequential organization is affected by proximity and continuity in frequency and time and the similarity in timbre. These factors are cooperative and competitive, and the relative importance of these factors for separation is a matter of debate and likely situation-dependent.

For the human auditory system, the competition and cooperation of simultaneous and sequential organization was explored by Bregman and Pinker [23]. They presented one well-known organization principle, the old-plus-new heuristic. It can be stated that "If any part of a sound can be plausibly interpreted as being a continuation of an earlier sound, then it should be.", which gives priority to sequential grouping. In this section, following their classical experiment, we give an example of how different principles compete and cooperate to control the separation of mixtures.

A. Methods and Stimuli

As shown in Fig. 13, for alternating tones A, B, and C, the simultaneous organization tends to integrate B and C into one source, which is accompanied by a series of pure tones A. However, the sequential grouping facilitates the integration of A and B and destroys the integration of the complex tone B-C. It can be thought that tones A and C compete to decide which one can be combined with B. In this experiment, the calculation of SI-SDRi is based on the separation of tones A vs. complex tones B-C, where source 1 is sequence A, and source 2 is complex tones B and C. For the other possible separation pattern, that source 1 is the alternating tones A and B, while source 2 is tone C, the results are similar and will not be repeated here.

In our experiment, four factors and their interactions are considered to analyze the dominant factors for fusion. The frequency of B-tones is fixed at 220 Hz, and tone duration is 100 ms. A-tones separate from B in frequency by 0 to 12 semitones in steps of 1 semitone and TRT between A and B is varied from 100 ms to 300 ms in steps of 20 ms. These two parameters are used to control the proximity in frequency and time for the strength of sequential grouping. On the contrary, the harmonicity and onset synchrony between tones B and C determine simultaneous organization. C tones are varied from 10 semitones to 14 semitones below B tones in steps of 0.5 semitones, when ΔF is -12 semitones means that C tones (110 Hz) and B tones (220 Hz) are harmonically related. The onset of C tones is varied from 0 to 50 ms behind B in steps of 5 ms to investigate the contribution of common onsets to fusion.

B. Results

The separation results as a function of the four factors are shown in Fig. 14. There is a total of 13*11 subgraphs, which represent the results under different ΔF between A and B (ΔF_{AB}) and different conditions of TRT. These two parameters control sequential grouping. In each subgraph, SI-SDRi (dB) is given as function of ΔF between B and C (ΔF_{BC}) and Δonset between tones B and C ($\Delta onset_{BC}$), which control simultaneous integration.

We start with the overall analysis of the effect of ΔF_{AB} and TRT for sequential grouping. For all subgraphs under different ΔF_{AB} and TRT, they show one generally that when ΔF_{AB} is larger and TRT is shorter, the separation between tones A and complex tones B-C is better. This demonstrates that the proximity principle in frequency and time for sequential separation is effective. When the tone interval in frequency and time is higher than the temporal coherence boundary, they tend to be separated.

Fig. 14 is divided into panels (a)–(d) according to separation performance. For panels (a) and (b), ΔF_{AB} is large enough ($\geq$6 semitones) for sequential organization to separate tones A and B according to the proximity principle. The presence of tones C also facilitates the grouping of tones C and B and the separation of tones A from tones B. In this condition, sequential and simultaneous organization are cooperative to separate tones A from the complex of tones B-C. As shown in the spectrogram (a1) and (b1), sources 1 and 2 are correctly separated.

For panel (c), ΔF_{AB} is less than 6 semitones and the time interval is less than 200 ms (TRT $\leq$ 200 ms). There is fierce competition between sequential and simultaneous grouping. The sequential organization here tends to combine tones A and B because $\Delta F_{AB} < 6$ semitones, while tones B and C are also affected by simultaneous organization. The spectrograms of three typical

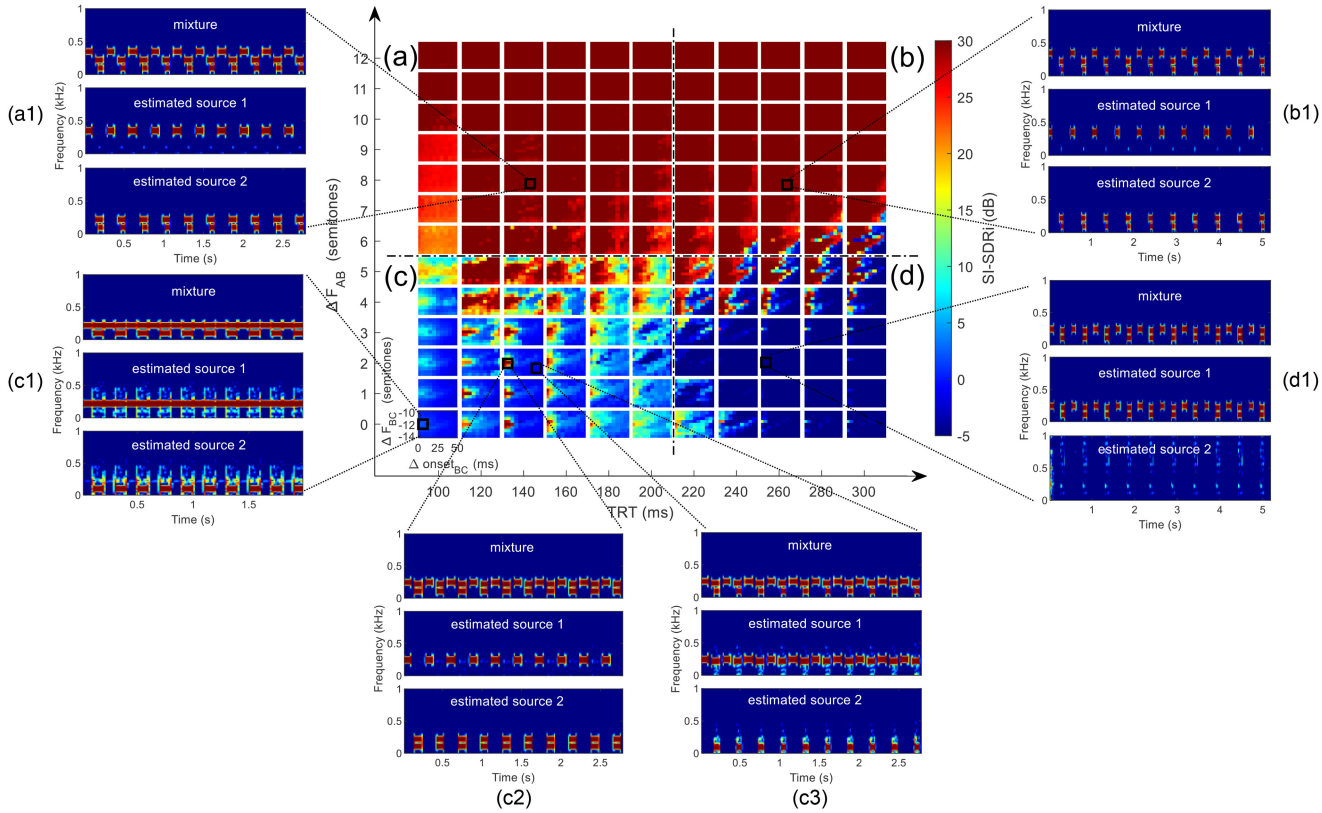


Fig. 14. For separation of source 1 (tones A) and source 2 (complex tones B-C), each subgraph represents SI-SDRi (dB) as a function of $\Delta\text{onset}_{\text{BC}}$ (ms) and $\Delta\text{F}_{\text{BC}}$ (semitones), both controlling simultaneous integration, while the overall axes present subpanels as a function of TRT (ms) and $\Delta\text{F}_{\text{AB}}$ (semitones), controlling sequential integration. (a)–(d) panels are divided according to the different separation performances. In addition, 6 typical cases (a1) (b1) (c1) (c2) (c3) (d1) are marked by black boxes, whose spectrograms of mixture and two sources estimated by the network are shown, respectively.

cases (c1) (c2) (c3) are shown in Fig. 14. For (c1), it is a special case that tones A and B are continuous (frequency of tones A and B is 220 Hz and TRT = 100 ms), resulting in a strong sequential grouping of tones A and B, which destroys the simultaneous grouping of tones B and C. For the other subgraphs of panel (c), if tones B and C are harmonic ($\Delta\text{F}_{\text{BC}} = -12$ semitones) and synchronous ($\Delta\text{onset}_{\text{BC}} = 0$ ms), they will integrate into one source regardless of the proximity between tones A and B, as visible in panel (c2). In this situation, it is plausible to believe that harmonicity and synchrony for simultaneous grouping are stronger than effects of sequential grouping. With the introduction of onset asynchrony or mistuning of harmonics, the force of simultaneous grouping between B and C weakens. As shown in panel (c3), tones B combine with A in sequence again, and tones C are separated.

For panel (d), especially for $\Delta\text{F}_{\text{AB}} \leq 3$ semitones, regardless of the relationship between tones B and C, the separation performance is very poor. As shown in panel (d1), three tones are grouped into one source. According to the proximity principle, for large TRTs, sequential organization will force the components to integrate into one stream, i.e., the separation process is dominated by sequential organization. In summary, the analysis shows that using ConvTasNet for segregation, principles for simultaneous and sequential separation compete and cooperate in dealing with segregation of acoustic scenes similar to the auditory system, and the relative importance of principles for separation depends on the specific situation.

VI. Grouping Based on Harmonicity in Complex Stimuli

In the above sections, simple artificial stimuli commonly used in psychoacoustic experiments are adopted to illustrate that Gestalt principles have been acquired by the network. We now explore whether these segregation principles generalize to more complex stimuli, such as speech.

Harmonicity is a prominent characteristic of the voiced parts of speech and it has been established in many psychoacoustic experiments to play a critical role in natural sound source separation [17]. In Section IV we have demonstrated the segregation of two complex tones with different F0s based on the harmonicity principle. We now explore the effectiveness of harmonicity in the separation of speech by destroying the harmonicity. McDermott *et al.* [53] used the STRAIGHT with sinusoidal modeling [54] to manipulate the harmonicity in speech and conducted psychoacoustic experiments to reveal the role of harmonicity for natural speech separation. The results showed that inharmonic speech was less intelligible for concurrent sentences, indicating that harmonicity contributes to auditory segregation.

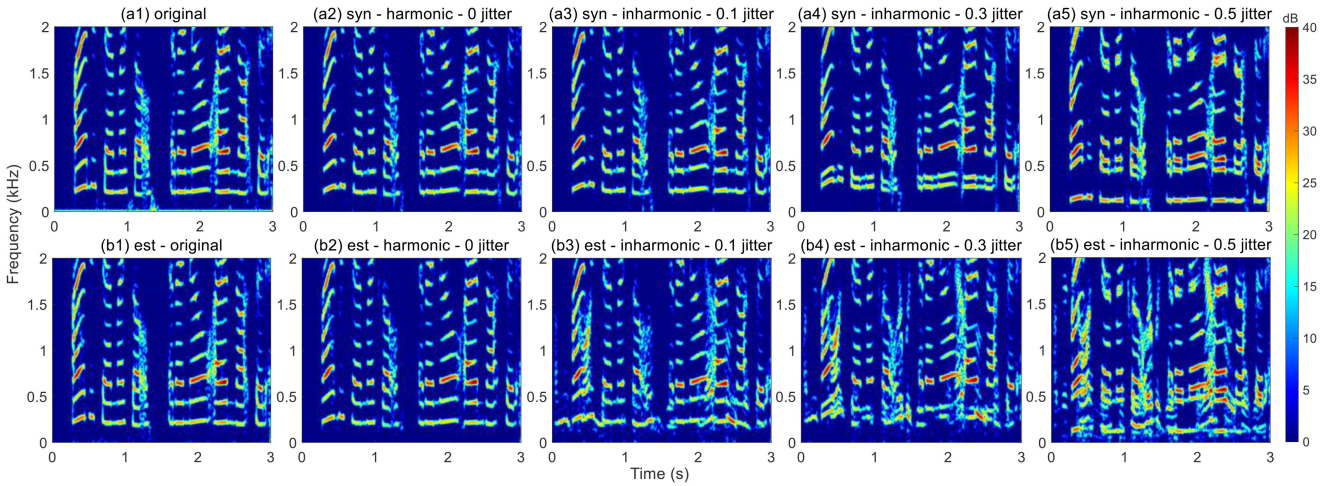


Fig. 15. The spectrogram of one sentence spoken by a female speaker, where (a1)–(a5) indicate the original speech sentence, synthetic harmonic speech and synthetic inharmonic speech jittered by 0.1, 0.3, 0.5 of the F0, and (b1)–(b5) indicate the corresponding estimates by the network.

A. Methods and Stimuli

STRAIGHT [55] is a speech analysis and synthesis vocoder. For a speech utterance, STRAIGHT estimates speech parameters of voiced excitation (the time-varying F0), unvoiced excitation (the time-varying spectral parameters of aperiodic components), and vocal tract filter (the time-varying spectral envelope). Then these estimated parameters can be manipulated to synthesize more altered speech signals. Here, STRAIGHT with sinusoidal modeling is used to generate speech with inharmonic carrier components while preserving other attributes. It models the voiced excitation as a sum of sinusoids, permitting frequency components to be manipulated individually.

In this experiment, the random jittering manipulation to each component is adopted rather than shifting by fixed frequencies to ensure that components no longer have any spectra regularity.

The first 30 harmonics of sources are randomly jittered individually. The n-th inharmonic carrier component f_n is generated by jittering the n-th harmonic with a random proportion of F0,

$$f_n = n\mathrm{F0} + c_n\mathrm{F0},$$

where c_n is a random value that follows a uniform distribution between $-c$ and c. The jitter magnitude (c) is used to control the degree of inharmonicity, which is fixed as 0.1, 0.2, 0.3, 0.4, and 0.5.

Taking one sentence spoken by a female speaker as an example, the spectrogram of the original speech sentence, synthetic harmonic speech, and synthetic inharmonic speech jittered by 0.1, 0.3, 0.5 are shown in Fig. 15(a1)–(a5), respectively. It can be seen that the synthetic harmonic speech is highly similar to the original speech, demonstrating the high accuracy of STRAIGHT for analysis and synthesis. For inharmonic speech, each component is randomly shifted upwards or downwards with a random proportion of F0, while the spectrotemporal envelope that conveys the information is preserved. In quiet, it sounds like harmonic speech accompanied by some whistle or some reverberation.

In this experiment, 1000 speech sentences are selected from the LibriSpeech dataset [37], which are not included in the network training dataset. For each sentence, 5 jitter magnitudes (c = 0.1, 0.2, 0.3, 0.4, 0.5) are explored, and 100 random jitter patterns for each jitter magnitude are created to avoid accidental harmonicity in a random jitter pattern. To create mixtures, two synthetic sentences with the same jitter degree are selected randomly and mixed with equivalent energy. The cross mixture of different jitter degrees is not tested, such as 0.1 jitter for one sentence and 0.2 jitter for another sentence. There is a total of $1000\times5\times100$ mixture clips, and each clip is 3 s in length with a 16 kHz sample frequency.

B. Results

Before the analysis of experiment results, the spectrograms of one example sentence spoken by a female speaker are shown in Fig. 15. When the example sentence and another sentence spoken by a male speaker are mixed correspondingly (the same degree of inharmonicity), the spectrograms of the network estimation are shown in (b1)–(b5). It is visually apparent that the example sentence is well separated and reconstructed for both original and synthetic harmonic cases in (b1) and (b2). However, when the harmonic components are perturbed by random jitters, even by 0.1 of the F0, the separated sentence is filled with other interferences and loses some necessary components. If frequency components are not subject to strict harmonic constraints, their assignment to appropriate sources is more difficult for the network and the separation performance decreases with the increase of the degree of inharmonicity.

Results of separating two concurrent sentences with different degrees of inharmonicity are given by SI-SDRi (dB) in Fig. 16. For mixtures of 1000 speech sentences from the LibriSpeech dataset (‘original’), the average SI-SDRi is 14.47 dB, which is comparable with the result of speech and speech separation reported in Section III. For synthetic harmonic mixtures, the average SI-SDRi is 12.46 dB. This reduction of about 2 dB is due to the accuracy of estimation and synthesis of STRAIGHT,

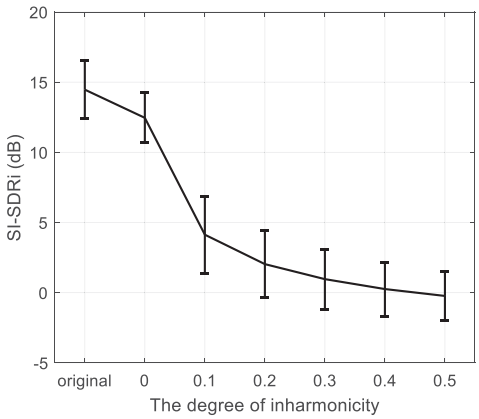


Fig. 16. SI-SDRi (dB) as a function of the degree of inharmonicity in the vowels of a synthetic sentence. 'Original' indicates the separation of the two original sentences, '0' indicates the separation of two synthetic harmonic sentences, and '0.1'–'0.5' indicates the separation of two synthetic inharmonic sentences with harmonics jittered by 0.1–0.5 of the F0. Error bars indicate the standard deviation of results for $1000\times5\times100$ mixture clips.

where some instantaneous components are not accurately synthesized. For the effect of harmonicity, once the harmonicity is destroyed, even jittered by 0.1 of the F0 will make the average separation performance drop sharply to 4.13 dB. As the degree of inharmonicity further increases, the separation performance decreases gently and tends to level off when the jitter magnitude is larger than 0.3 ($c > 0.3$). The performances in our network are consistent with that in psychoacoustic experiments obtained by McDermott *et al.* [53], which illustrated that the intelligibility of concurrent words or sentences decreased with the degree of inharmonicity for the human auditory system.

In summary, for concurrent sentences, inharmonicity hinders the grouping of frequency components and the separation performance decreases with the degree of inharmonicity increases. This suggests that the network uses harmonicity principles to separate speech. The harmonicity principle is still effective in the separation of complex sound sources.

VII. Discussion

Two general approaches have been followed to solve sound source separation problems: one is to separate sources through the imitation of the auditory system, while the other is to base it solely on statistical signal processing.

The first approach develops a model with biological rational based on knowledge from psychoacoustics and auditory neuroscience (CASA models) [2]. Conceptually, CASA models operate as a two-stage process: segmentation and grouping. Segmentation is based on the representation of sound attributes. Hu and Wang [4] estimated attributes (pitch and AM) through the imitation of auditory peripheral and mid-level processing. Recently, more auditory central processes have appeared in CASA models. Elhilali *et al.* [56] mimicked human cortical processing to segregate auditory objects, which mapped the acoustic waveform into a 4-D cortical representation. Elhilali *et al.* [12] learned spectro-temporal representations through a stochastic neural network with two layers, including a local analysis layer and a long-range analysis layer to mimic the simultaneous and sequential organization in the auditory system respectively.

After obtaining the effective representation of sound attributes, CASA models tend to group segments on the basis of Gestalt principles, such as the proximity in pitch explored by Hu and Wang [4], dynamic similarity reflected by an integrative and clustering stage [56], and temporal coherence via Hebbian learning [12]. One or few specific grouping principles, rather than all principles, are implemented in traditional CASA models and dominate the process of separation. If CASA models are tested with the simple stimuli in our study, they are likely to have similar behavior characteristics with humans in some principles but not all aspects.

Traditional CASA models separate sound sources by carefully modeling the auditory system, while the current understanding of auditory neuroscience is not sufficient to develop a system as intelligent as humans. The opportunities to learn principles through task optimization are lost in those unsupervised models, which makes them particularly effective for simple stimuli and cannot be generalized to natural sources in complex scenes.

The second approach treats source separation as a supervised learning problem. In recent years, these statistical deep network models have achieved excellent performance in natural source separation. However, few attempts have been made to investigate the biological plausibility of this kind of network because these models do not appear to mimic the auditory system in a particular way and network weights are hard to interpret. In this paper, we demonstrate that similar separation mechanisms emerge in our statistical network as they are present in human hearing. This means that without accurate biological modeling, a network that follows Gestalt rules can be obtained.

Is the ultimate destination of deep learning to be able to spontaneously learn the same optimization criteria like human beings? Francl and McDermott [11] showed that for localization, trained networks can spontaneously operate similarly to human spatial hearing. In our study, the trained network has also been shown to behave similarly to human hearing – in this case the more complex auditory scene analysis. We believe that observing the emergence of (segregation) mechanisms purely on the basis of statistical signal processing has a profound influence on the study of deep learning and auditory neuroscience.

The trained network is not limited to some specific sound sources but depends on generalized primitive grouping principles. The underlying generalization suggests a general source separation network that can adapt to all scenes and achieve selective hearing like the human auditory system. The work also provides a new perspective on network interpretation: the underlying mechanisms are explored through Gestalt experiments following those developed in many years of auditory research, which goes beyond the visualization of features or filter activation and can be used to probe specific hypotheses while building on a wealth of previous experience. The hypothesis testing helps explain the "black box" of the network and in turn guides further network optimization.

On the other hand, the emergence of separation mechanisms through network learning can also help our understanding of the processes in the human auditory system. As an ideal observer,

the network can nonetheless test a lot of stimuli, which could be instructive for future psychoacoustic experiments.

VIII. Conclusion

In this study, a convolutional deep neural network, ConvTasNet, is developed to separate arbitrary sounds in the time domain, including speech, music, and environmental sounds. The SI-SDRi of the best-performing network is 11.70 dB, which is comparable or even surpasses the result of IRMs. This demonstrates that our network has an excellent ability to separate natural complex sound sources.

For this network, that is capable of solving actual separation problems, the underlying separation mechanisms are investigated. At first, the method of Gestalt psychologists is adopted, where simple stimuli are used to explore principles one by one on simultaneous and sequential organization. There are fundamental differences between the training dataset (natural sources) and the highly specific and abstracted artificial stimuli made of tones. These differences make sure that only when networks generalize segregation principles, the separation performance of untrained artificial stimuli may follow that of human auditory scene analysis. Then, speech stimuli are generated with inharmonic carrier components while preserving other attributes to explore whether the harmonicity principle can be generalized to complex stimuli.

To our knowledge, it is the first demonstration that Gestalt principles underlying human auditory scene analysis are learned by supervised deep learning from unrelated sound sources with a completely statistical model that does not have any particular auditory-related process. The experiments probing simultaneous organization demonstrate that harmonicity, onset asynchrony, and coherent AM and FM assist the segregation. For sequential grouping, proximity in time and frequency is in a consistent manner with the emergence of a temporal coherence boundary like in the auditory system. A good continuation in frequency also exerts a strong force to integrate components into sequences. The similarity in timbre, as studied with varying spectral centroid, contributes to separation beyond the effects of fundamental frequency and intensity. These principles for simultaneous and sequential organization are shown to be cooperative and competitive, and the relative importance of these principles for separation is situation dependent. In addition, the experiment of concurrent sentence separation illustrates that the harmonicity principle is still effective in the separation of complex sound sources.

In summary, without prior knowledge about auditory scene analysis principles imparted on the network, it learns separation mechanisms similar to those in the human auditory system, which provides a new perspective for the problem of auditory scene analysis. Since ConvTasNet is a purely statistical model aiming to optimally segregate sound sources, results suggest that the mechanisms developed in the auditory system over many years have evolved for optimal segregation based on statistical characteristics of the acoustical signal.

Our study is the first step for exploring auditory-like mechanisms learned by deep networks. In the future, the comparison of network results and psychoacoustic experiments may extend to other experiments and objective measures. In addition, how the networks' specific structure contributes to performance over data statistics is also worth exploring. It is not yet clear whether other kinds of networks can learn similar separation mechanisms from unrelated natural sounds.

References

[1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1990.

[2] D. Wang and G. J. Brown, "Fundamentals of computational auditory scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. J. Brown, Eds., Hoboken, NJ, USA: Wiley, 2006, pp. 1–37.

[3] D. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, May 1999.

[4] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[5] G. Hu and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 2, pp. 396–405, Feb. 2007.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[7] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 189–198, Jan. 2019.

[8] Y. Liu and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[10] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 46–50.

[11] A. Francl and J. H. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nature Hum. Behav.*, vol. 6, no. 1, pp. 111–133, 2022.

[12] D. Chakrabarty and M. Elhilali, "A Gestalt inference model for auditory scene segregation," *PLoS Comput. Biol.*, vol. 15, no. 1, pp. 1–33, 2019.

[13] H. Li, K. Chen, and B. U. Seeber, "Auditory filterbanks benefit universal sound source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 181–185.

[14] H. Li, K. Chen, R. Li, J. Liu, B. Wan, and B. Zhou, "Auditory-like simultaneous separation mechanisms spontaneously learned by a deep source separation network," *Appl. Acoust.*, vol. 188, 2022, Art. no. 108591.

[15] M. Wertheimer, "Untersuchungen zur Lehre von der Gestalt. II," *Psychol. Forsch.*, vol. 4, no. 1, pp. 301–350, Jan. 1923.

[16] C. J. Darwin and R. P. Carlyon, "Auditory grouping," in *Hearing*, 2nd ed., B. C. J. Moore, Ed., San Diego, CA, USA: Academic, 1995, pp. 387–424.

[17] C. Micheyl and A. J. Oxenham, "Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings," *Hear. Res.*, vol. 266, no. 1–2, pp. 36–51, 2010.

[18] J. F. Culling and C. J. Darwin, "Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating," *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1559–1569, 1994.

[19] G. J. Sandell and C. J. Darwin, "Recognition of concurrently-sounding musical instruments with different fundamental frequencies," *J. Acoust. Soc. Amer.*, vol. 100, no. 4, Oct. 1996, Art. no. 2683.

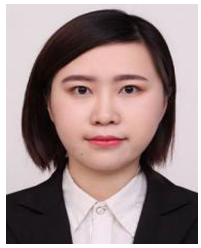
[20] C. J. Darwin, "Pitch and auditory grouping," in *Pitch: Neural Coding and Perception*. New York, NY, USA: Springer, 2005, pp. 278–305.

[21] N. Grimault, S. P. Bacon, and C. Micheyl, "Auditory stream segregation on the basis of amplitude-modulation rate," *J. Acoust. Soc. Amer.*, vol. 111, no. 3, pp. 1340–1348, 2002.

[22] C. J. Darwin and V. Ciocca, "Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component," *J. Acoust. Soc. Amer.*, vol. 91, no. 6, pp. 3381–3390, 1992.

[23] A. S. Bregman and S. Pinker, "Auditory streaming and the building of timbre," *Can. J. Psychol.*, vol. 32, no. 1, pp. 19–31, 1978.

[24] J. H. Lee and L. E. Humes, "Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background," *J. Acoust. Soc. Amer.*, vol. 132, no. 3, pp. 1700–1717, 2012.
[25] A. S. Bregman, J. Abramson, P. Doehring, and C. J. Darwin, "Spectral integration based on common amplitude modulation," *Percep. Psychophys.*, vol. 37, no. 5, pp. 483–493, 1985.
[26] R. P. Carlyon, "The psychophysics of concurrent sound segregation," *Philos. Trans. Roy. Soc. London. Ser. B: Biol. Sci.*, vol. 336, no. 1278, pp. 347–355, Jun. 1992.
[27] R. P. Carlyon, "Discriminating between coherent and incoherent frequency modulation of complex tones," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 329–340, Jan. 1991.
[28] M. H. Chalikia and A. S. Bregman, "The perceptual segregation of simultaneous vowels with harmonic, shifted, or random components," *Percep. Psychophys.*, vol. 53, no. 2, pp. 125–133, 1993.
[29] N. Itatani and G. M. Klump, "Animal models for auditory streaming," *Philos. Trans. Roy. Soc. B: Biol. Sci.*, vol. 372, no. 1714, 2017, Art. no. 20160112.
[30] L. P. A. S. van Noorden, "Temporal coherence in the perception of tone sequences," Ph.D. dissertation, Inst. Perception Res., Technische Hogeschool Eindhoven, Eindhoven, 1975.
[31] P. G. Singh, "Perceptual organization of complex-tone sequences: A tradeoff between pitch and timbre," *J. Acoust. Soc. Amer.*, vol. 82, no. 3, pp. 886–899, 1987.
[32] R. Cusack and B. Roberts, "Effects of differences in timbre on sequential grouping," *Percep. Psychophys.*, vol. 62, no. 5, pp. 1112–1120, 2000.
[33] D. L. Wessel, "Timbre space as a musical control structure," *Comput. Music J.*, vol. 3, no. 2, pp. 45–52, 1979.
[34] I. Kavalerov *et al.*, "Universal sound separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 175–179.
[35] H. Li, K. Chen, and B. U. Seeber, "ConvTasNet-based anomalous noise separation for intelligent noise monitoring," in *Proc. INTERNOISE*, 2021, pp. 2044–2051.
[36] "The BBC sound effects library," [Online]. Available: http://bbcsfx.acropolis.org.uk/
[37] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.
[38] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, *arXiv:1510.08484*.
[39] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - Half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.
[40] D. Yu, M. Kolbaek, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 241–245.
[41] M. Pariente *et al.*, "Asteroid: The pytorch-based audio source separation toolkit for researchers," in *Proc. INTERSPEECH*, 2020, pp. 2637–2641.
[42] A. Van Den Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
[43] G. Wichern *et al.*, "WHAM!: Extending speech separation to noisy environments," in *Proc. INTERSPEECH*, 2019, pp. 1368–1372.
[44] A. J. Oxenham, "Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants," *Trends Amplification*, vol. 12, no. 4, pp. 316–331, 2008.
[45] A. S. Bregman and G. L. Dannenbring, "The effect of continuity on auditory stream segregation," *Percep. Psychophys.*, vol. 13, no. 2, pp. 308–312, 1973.
[46] R. P. Carlyon and H. E. Gockel, "Effects of harmonicity and regularity on the perception of sound sources," in *Auditory Perception of Sound Sources*, Boston, MA, USA: Springer, 2008, pp. 191–213.
[47] N. Jacoby, E. A. Undurraga, M. J. McPherson, J. Valdés, T. Ossandón, and J. H. McDermott, "Universal and non-universal features of musical pitch perception revealed by singing," *Curr. Biol.*, vol. 29, no. 19, pp. 3229–3243, Oct. 2019.
[48] E. Terhardt, "Calculating virtual pitch," *Hear. Res.*, vol. 1, pp. 155–182, 1979.
[49] R. Plomp and W. J. M. Levelt, "Tonal consonance and critical bandwidth," *J. Acoust. Soc. Amer.*, vol. 38, no. 4, pp. 548–560, Oct. 1965.
[50] W. A. Sethares, "Local consonance and the relationship between timbre and scale," *J. Acoust. Soc. Amer.*, vol. 94, no. 3, pp. 1218–1228, Sep. 1993.
[51] G. L. Dannenbring and A. S. Bregman, "Streaming vs. fusion of sinusoidal components of complex tones," *Percep. Psychophys.*, vol. 24, no. 4, pp. 369–376, 1978.
[52] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed., New York, NY, USA: Springer, 2007, pp. 247–253.
[53] S. Popham, D. Boebinger, D. P. W. Ellis, H. Kawahara, and J. H. McDermott, "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," *Nature Commun.*, vol. 9, no. 2122, pp. 1–13, 2018.
[54] J. H. McDermott, D. P. W. Ellis, and H. Kawahara, "Inharmonic speech: A tool for the study of speech perception and separation," in *Proc. SAPA-SCALE*, 2012, pp. 114–117.
[55] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *SADHANA*, vol. 36, no. 5, pp. 713–727, 2011.
[56] M. Elhilali and S. A. Shamma, "A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation," *J. Acoust. Soc. Amer.*, vol. 124, no. 6, pp. 3751–3771, 2008.



Han Li received the B.Eng. degree in environmental engineering in 2015 from Northwestern Polytechnical University, Xi'an, China, where she is currently working toward the Ph.D. degree in acoustics. From 2018 to 2020, she was the recipient of the two -year Ph.D. scholarship from the China Scholarship Council to do research with the Technical University of Munich, Munich, Germany. Her research interests include source separation, sound event detection, and auditory perception.

Kean Chen received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 1992. He is currently a Professor with the School of Marine Science and Technology, Northwestern Polytechnical University. His research interests include active control of noise and vibration, auditory perception and its applications. Prof. Chen is a member of the Acoustical Society of China, a Member of Chinese Society for Vibration Engineering, the President of the Acoustical Society of Shaanxi Province, the Vice Director of the Noise Control Chapter of Chinese Society for Vibration Engineering, the Vice Director of the Environmental Acoustics Chapter of the Acoustical Society of China, a Committee Member of National Acoustics Standardization Committee, and an Editorial Board Member of the *Journal of Vibration Engineering*, *Noise and Vibration Control*, and *Journal of Marine Engineering*.

Bernhard U. Seeber received the Dipl.-Ing. degree in electrical engineering and information technology and the Dr.-Ing. degree (with distinction) from the Technical University of Munich (TUM), Munich, Germany, in 1999 and 2003, respectively. He was a Postdoc with the Department of Psychology, University of California, Berkeley, Berkeley, CA, USA. In 2007, he joined the MRC Institute of Hearing Research, Nottingham, U.K., to lead the Spatial Hearing Laboratory. Since 2012, he is the Head of the Audio Information Processing Lab and a Professor with the Department of Electrical and Computer Engineering, TUM. His research interests include signal processing for hearing aids and cochlear implants, on virtual acoustics, spatial hearing, auditory modeling and acoustic nondestructive testing. Prof. Seeber is a Member of the German Acoustical Society (DEGA), Association for Electrical, Electronic and Information Technologies, Acoustical Society of America, Association for Research in Audiology, and Bernstein Network for Computational Neuroscience. He heads the Technical Committee on hearing acoustics in the Society for Information Technology (ITG/VDE) and was a Member of the Executive Board of the DEGA from 2016 to 2022. He was the recipient of the Lothar-Cremer Award of the DEGA, Doctoral Thesis Award of the ITG, and ITG Publication Award.