Contents lists available at ScienceDirect

Journal of Otology

journal homepage: www.journals.elsevier.com/journal-of-otology/

Review

Speech perception in noise: Masking and unmasking

Xianhui Wang*, Li Xu

Communication Sciences and Disorders, Ohio University, Athens, OH, 45701, USA

ARTICLE INFO

Article history:
Received 9 October 2020
Received in revised form
3 December 2020
Accepted 6 December 2020

Keywords:
Speech perception
Auditory masking
Speech unmasking
Cocktail party problems
Mandarin Chinese

ABSTRACT

Speech perception is essential for daily communication. Background noise or concurrent talkers, on the other hand, can make it challenging for listeners to track the target speech (i.e., cocktail party problem). The present study reviews and compares existing findings on speech perception and unmasking in cocktail party listening environments in English and Mandarin Chinese. The review starts with an introduction section followed by related concepts of auditory masking. The next two sections review factors that release speech perception from masking in English and Mandarin Chinese, respectively. The last section presents an overall summary of the findings with comparisons between the two languages. Future research directions with respect to the difference in literature on the reviewed topic between the two languages are also discussed.

Contents

1. Introduction 109
2. Auditory masking 110
3. Speech unmasking in English 111
    3.1. Low-level unmasking factors 111
        3.1.1. Spatial separation 111
        3.1.2. Speaker voice difference 112
    3.2. High-level unmasking factors 113
        3.2.1. Linguistic variables 113
        3.2.2. Attention and cognitive factors 114
4. Speech unmasking in Mandarin Chinese 115
    4.1. Spatial separation 115
    4.2. Speaker voice difference 115
    4.3. Linguistic variables 115
    4.4. F0 contours 115
    4.5. Target priming 116
5. Summary 116
Funding 116
Declaration of competing interest 116
Acknowledgments 116
References 116

1. Introduction

Speech carries linguistic information that is important for human communication and social activities. Essentially, it is an

* Corresponding author.
E-mail addresses: xw659217@ohio.edu (X. Wang), xul@ohio.edu (L. Xu).

Peer review under responsibility of PLA General Hospital Department of Otolaryngology Head and Neck Surgery.

https://doi.org/10.1016/j.joto.2020.12.001

acoustic signal produced by an “energy source”: exhalation of the lungs and followed by fine modulations by the vocal tract and articulators (i.e., the source-filter model, Fant, 1960). Although speech is relatively robust to noise interference (see Diehl, 2008 and Bronkhorst, 2015 for details), speech perception can be challenging when the interfering noise is also speech (i.e., speech-on-speech masking). Cherry (1953) first used the term “*Cocktail Party Problem*” to describe the circumstance where a listener selectively attends to one specific speech among other concurrent speech (see Fig. 1). Since then, more and more hearing scientists started to explore the sensory limits of ears in parsing auditory scenes.

So far, research conducted in English has revealed several factors that facilitate “unmasking” of the speech target (i.e., improving speech perception) in cocktail-party listening environments. These factors include the target/masker spatial separation and the difference in target/masker voice characteristics. The magnitude of the unmasking benefits due to these factors were also found to be consistently large across a number of reports. Interestingly, recent studies in Mandarin Chinese, which is a tonal language, replicated these English experiments and observed different patterns: the magnitude of unmasking benefits for spatial separation seemed to be smaller, but larger for voice characteristics compared to English (e.g., Wu et al., 2005; 2011; Zhang et al., 2020; Chen et al., 2020).

Despite the methodological differences between the Mandarin Chinese studies and English studies, the results have led to a possibility that the linguistic differences between Mandarin Chinese and English can carry over to the difference in the degree to which both languages benefit from spatial separation and difference in voice characteristics. These linguistic differences between Mandarin Chinese and English may include but are not limited to (1) the quantity of voiced consonants is fewer in Mandarin Chinese (Kang, 1998), (2) the pitch contour carries lexical meaning in Mandarin Chinese but not in English (Liang, 1963), and (3) most Mandarin Chinese words are multimorphemic, whereas most English words are monomorphemic (Wu et al., 2011). Nevertheless, the topic of speech unmasking in cocktail-party situations has been under-researched in Mandarin Chinese compared to English.

In order to better understand the differences in unmasking benefits in Mandarin Chinese and English and the underlying mechanisms, this review first presents a series of concepts regarding auditory masking which are important for speech perception in cocktail party situations. The following sections summarize findings on unmasking factors in speech-on-speech masking experiments in English and Mandarin Chinese, respectively. The last section provides an overall summary and comparison of the relevant findings between the two languages. Future research directions are then discussed.

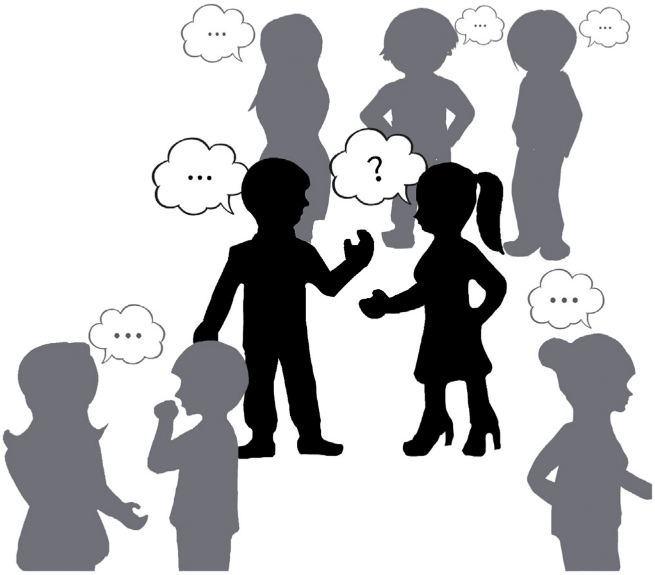

Fig. 1. Illustration of the cocktail party problems. For example, the woman in black experienced difficulties in understanding the speech from the man in black due to the presence of background talkers.

2. Auditory masking

People often have trouble in understanding speech when the communication takes place in adverse listening environments. From the hum of a car engine to the echo in a reverberant room, noise can be any kind of sound that is “unwanted” to the listener and it usually compromises speech perception. This deficit in perceiving a sound of interest (for example, speech) is a consequence of *Auditory Masking*, defined as the process by which the threshold of hearing one sound is raised by the presence of another (Moore, 2012). Consequently, increasing the intensity of one sound relative to other confounding sounds should benefit its perception. Indeed, the primary analysis of a sound mixture at the cochlea (Fletcher, 1940) appears to give credit to spectrotemporal (i.e., frequency-time) units with higher energy (Brown and Cooke, 1994; Wang and Brown, 1999). Specifically, when a sound mixture of source A and B is decomposed into different frequency constituents on the basilar membrane at an instantaneous time point, the source with a higher energy at that specific time-frequency point will dominate the neural representation in the auditory nerve (see Fig. 2).

However, this “winner-takes-all” mechanism (i.e., ideal binary masks, Brown and Cooke, 1994) could be a problem when attributes of the unwanted sound dominate a sufficient number of spectrotemporal units along a long time period. In this case, the corresponding neural representation will be occupied by the unwanted sound, rendering the target sound ‘inaudible’ to auditory nerves. This kind of masking arising from the overwhelming neural representation of masking sounds is referred to as *Energetic Masking* (EM). Although it has been widely accepted that EM predominantly occurs at the auditory periphery, it was argued that such definition based on the level of processing may not be accurate (Culling and Stone, 2017). For example, some processes that are able to reduce EM such as binaural processing (Durlach 2006; Hirsh 1948) can occur beyond the peripheral level.

Apart from EM, another type of masking that mainly occurs at the auditory periphery level is *Modulation Masking* (MM, Stone et al., 2012; Stone and Canavan, 2016; Culling and Stone, 2017). MM derives from the interaction of target and masker intrinsic modulation. Importantly, the random noise commonly used in speech perception in noise experiments that was thought to produce pure EM, in fact, has intrinsic envelope fluctuations (Stone et al., 2012; Stone and Canavan, 2016). The fluctuation in the masker’s envelope also obscures that of the target, so the total masking effect using a random noise should be accounted for by both EM and MM, instead of EM alone. By removing the intrinsic modulation fluctuation from the random noise, Stone et al. (2012) and Stone and Canavan (2016) revealed that “dip-listening” (i.e., listeners’ ability to glimpse a target signal through the spectrotemporal envelope fluctuation of masking noise) was mainly dependent on the release from MM rather than from EM.

On the other hand, sounds with relatively sparse spectrotemporal distribution can also disturb auditory perception. For example, only two to three simultaneous talkers in the background can be confusing enough for listeners to tell apart, even when the signal-to-noise ratio (SNR) is not challenging (Brungart et al., 2001).

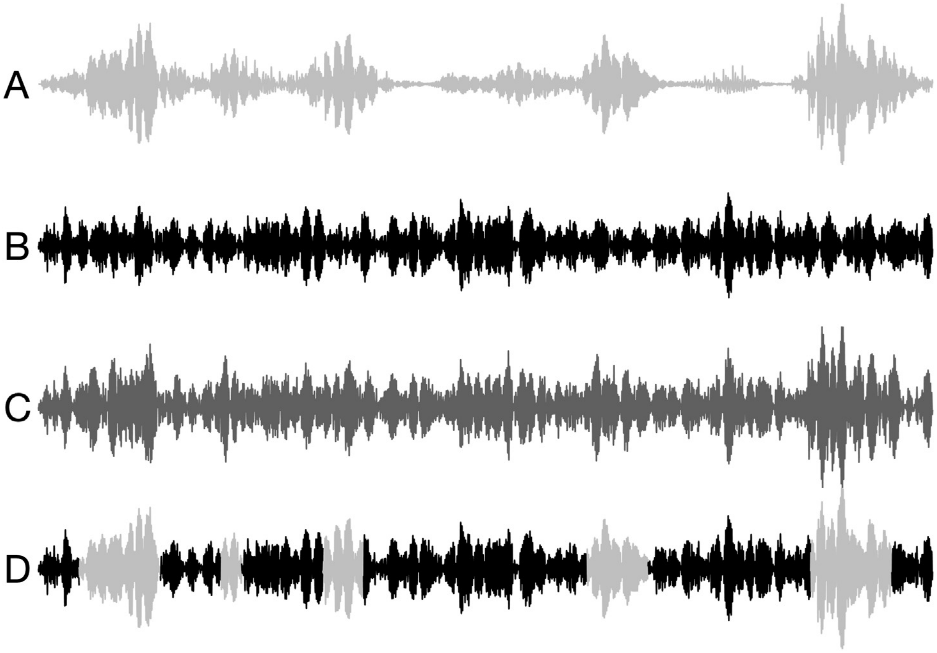

Fig. 2. Illustration of the "winner-takes-all" mechanism for energetic masking. A: Waveform of the first 884 msec of a Mandarin AzBio sentence (List F Sentence 15) (Xu et al., 2020) passed through a bandpass filter between 801 and 925 Hz. This band was approximately 1 ERB wide (Stone et al., 2012). B: Waveform of an 884-msec long speech-spectrum-shaped noise passed through the same bandpass filter. The amplitude of the noise was adjusted so that the root-mean-squared value was equal to that of the speech signal, thus the signal-to-noise ratio was 0 dB. C: Speech in noise (i.e., A + B). D: The "winner-takes-all" result. A time bin was set at 20 ms (Brungart et al., 2006). For all time bins in which the speech amplitude was greater than the noise amplitude, the speech waveform was plotted in light gray. Otherwise, the noise waveform was plotted in black.

Auditory researchers used "*Perceptual Masking*" (Carhart et al., 1968) or "*Informational Masking*" (IM, Brungart, 2001) to refer to the masking due to a group of central factors (e.g., attention or linguistic factors). In other words, IM occurs when both the target and masker are audible but not distinguishable and when the masker information is misattributed to the target (Carlile and Corkhill, 2015).

In a cocktail party situation, speech-on-speech masking may involve all these types of masking but to different extents. For example, when there are only a few background talkers, masking speech may remain intelligible with little spectrotemporal energy overlap and modulation interaction with the target. In this case, the difficulty in understanding a target speech primarily derives from IM. In contrast, if the number of background talkers is larger, EM and MM may account for a larger proportion of the total masking due to the increased spectrotemporal overlap and modulation disruption on the target, whereas IM may play a minor role as linguistic information of masking sounds become obscure (Hoen et al., 2007).

3. Speech unmasking in English

The perceptual outcome in a cocktail party problem is a result of the competition between masking and unmasking. Factors contributing to EM, MM, and IM make a cocktail party situation more challenging, whereas factors assisting auditory processing can unmask the target speech. The benefit provided by these unmasking factors can be quantified by measuring the improvement in behavioral performance relative to when unmasking factors are unavailable (i.e., masking release).

Some unmasking factors are labeled "low-level" while others are "high-level" (e.g., Mattys et al., 2012). This distinction is made based on the level of auditory processing that the unmasking process involves. Specifically, unmasking factors are considered as low-level when they mainly reduce energetic components of masking or benefit preattentional processes (or auditory object formation, Shinn-Cunningham, 2008) such as local grouping and streaming. In contrast, "high-level" unmasking factors facilitate central processes such as selective attention (i.e., auditory object selection, Shinn-Cunningham, 2008) and/or linguistic competition. The present section discusses both low- and high-level unmasking factors in cocktail-party situations where the difficulty in speech perception is mainly governed by IM.

3.1. Low-level unmasking factors

3.1.1. Spatial separation

When two sounds are presented to human ears from different locations, the task of attending to one sound becomes easier compared to when they are co-located. The improvement in auditory perception due to target/masker spatial separation is called spatial release from masking (SRM).

From a masking point of view, SRM reflects release from both EM and IM (e.g., Ihlefeld and Shinn-Cunningham, 2008). The energetic component can be reduced by processes such as binaural unmasking (e.g., Bronkhorst, 2000; Culling et al., 2004; Kidd et al., 2010; Marrone et al., 2008) and better-ear listening (e.g., Edmonds and Culling, 2006; Culling and Mansell, 2013; Culling and Stone, 2017). In a cocktail party situation, spatial separation mainly benefits speech perception by reducing IM (cf. Kidd et al., 2008). Listeners show large SRM in auditory perception tasks where the masker is highly similar to the target in linguistic aspects (e.g., Freyman et al., 2001; Kidd et al., 2010; Swaminathan et al., 2015; Viswanathan et al., 2016; Calandruccio et al., 2017) and in voice characteristics (e.g., Best et al., 2012; Zekveld et al., 2014; Xia et al., 2015; Kidd et al., 2016; Rennies et al., 2019), where the attention is distracted (e.g., Best et al., 2006; Kopco, 2009; see a review by Sussman, 2017), or where the cognitive load is high (e.g., Zekveld et al., 2014; Andéol et al., 2017; Xia et al., 2015). In contrast, when a modulated noise that produces barely any IM is used as the masker, listeners usually benefit little from spatial separation (e.g., Freyman et al., 1999 and 2001). Therefore, the improvement in SRM relative to the modulated noise condition often serves as an estimation of IM in masking experiments (Schneider et al., 2007).

The benefit of spatial separation has been found relatively large across a large number of studies. Freyman et al. (1999) measured listeners' identification of nonsense sentences spoken by a female through two loudspeakers, one from the front (0° azimuth) and another from 60° azimuth to the right. The masker was either speech-spectrum-shaped noise (SSN) or a female talker. In the co-located condition, the target-masker stimuli were presented from both loudspeakers without time delay. In the spatial-separated condition, the masker sound in the right loudspeaker slightly led the masker sound from the front in time by 4 ms. This time delay was the key to elicit the perceived spatial separation between the target and masker because of the precedence effect which refers to the phenomenon that the auditory system fuses two separate sounds into one when one sound lags the other by a short time of delay (Zurek, 1987). As a result, the apparent image of the masker shifted away from the target and toward the temporally leading loudspeaker (i.e., right), whereas the target remains at the center in the co-located condition. The recognition score in percentage correct was measured at multiple SNRs (i.e., −12, −8, −4, and 0 dB) and SRM defined as the difference in dB required for reaching the same performance level was calculated. For a female masker, SRM was approximately 14 dB (at 70% correct), compared with 8 dB obtained for SSN, yielding a release from IM of about 6 dB. These results were replicated in later studies by Freyman et al. (2001, 2004).

Many other studies using different speech materials have also reported a large release from IM due to spatial separation. For example, several studies using Coordinate Response Measure (CRM) in which the sentences are of the form "Ready [call sign] go to [color] [number] now." observed an SRM of >10 dB for intelligible speech maskers (mostly two-talker babble) but a much smaller SRM for SSN masker (e.g., Andéol et al., 2017; Arbogast et al., 2002; Gallun et al., 2013; Kidd et al., 2014). The large magnitude of SRM due to reduced IM was also consistent with studies in which the effect of head-shadow was accounted for (e.g., Marrone et al., 2008; Best et al., 2012; Swaminathan et al., 2015).

In studies attempting to isolate EM or IM, it became clear that SRM can be largely attributed to release from IM. For example, Arbogast et al. (2002) processed two speech sources into acoustically mutually exclusive frequency channels using a tone vocoder (see Xu, 2016; Xu et al., 2020) and created two types of masker: "different-band speech" (DBS) and "different-band noise" (DBN). The DBS consisted of intelligible speech in narrow frequency bands that did not contain target speech, whereas the DBN consisted of equally narrow bands of noise (unintelligible) in the bands that did not contain target speech. In both types of maskers, EM was minimized to the same amount by eliminating the spectral overlap. Thus, the difference in performance of the DBS and DBN conditions could be attributed to the difference in intelligibility. The observed SRM was 18 dB for the DBS masker, in comparison to <10 dB for the DBN masker. This difference in SRM reflected the amount of IM.

In the signal-processing technique proposed by Brungart et al. (2006), "ideal time-frequency segregation" (ITFS), speech sources were divided into narrow frequency channels with each channel then being subdivided into brief time intervals. The resulting time-by-frequency (T-F) matrix illustrates values representing energy contained in each T-F unit. In this procedure, units with the target-to-masker energy beyond a certain value (determined based on knowledge of the stimulus) were preserved and those below that value were discarded. In other words, this procedure represents an "ideal" listener who performs perfect local segregation processes (see Shinn-Cunningham et al., 2017 for details), therefore, the output of this procedure contained little IM. Kidd et al. (2016) compared SRM between ITFS processed (i.e., glimpsed) and unprocessed stimuli for two- or four-talker masker. The SRM was found much larger for unprocessed stimuli (19.2 dB and 12.4 dB for two/four-talker masker, respectively) than for glimpsed stimuli (less than 2 dB for both maskers). Because ITFS removes most of IM, this result was consistent with the view that spatial separation benefits speech perception mainly by reducing the IM component in speech-on-speech masking conditions. Similar results were also reported by Rennies et al. (2019).

3.1.2. *Speaker voice difference*

The difference in fundamental frequency (F0) can be used by listeners to distinguish between speech streams from different sources (e.g., different talkers). Because F0 also contributes to the perception of pitch, the release from masking due to F0 difference is likely to involve higher-level processing. Improved perception performance in speech-on-speech masking experiments due to the target/masker voice difference has been well documented in many studies (e.g., Brungart 2001; Brungart et al., 2001; Cullington and Zeng, 2008; Brown et al., 2010). In general, the more similar target/masker voices are, the worse the speech perception performance becomes compared to when target/masker voices differ greatly (such as different sexes).

Brungart (2001) presented listeners with two competing phrases spoken in the same-sex, different-sex, or same voice and asked them to identify the target phrase. Listeners scored 20 percentage points higher when the masker phrase was of different sex rather than of the same sex as the target phrase. The lowest performance was observed when the target and masker were spoken in the same voice. A similar pattern was observed in a follow-up study by Brungart et al. (2001) where the number of concurrent talkers was increased to four. Consistently, several later studies also reported a large masking release due to dissimilar target/masker sex (Noble and Perrett, 2002; Brungart and Simpson, 2002; Allen et al., 2008; Brungart et al., 2001, 2009; Cullington and Zeng 2008; Zekvel et al., 2014; Xia et al., 2015; Kidd et al., 2016, 2019; Rennies et al., 2019; Zhang et al., 2020).

Zekvel et al. (2014) used meaningful and semantically neutral sentences as stimuli and observed a 4-dB benefit in SRT by using different sexes for the masker and target speakers. They also observed that spatial separation benefits were much larger for same-sex than different-sex maskers, indicating that the composition of IM was higher when the target/masker sex was the same. A possible explanation was that the listeners' cognitive processing load was higher (associated with a larger pupil response) when the masker and target voices were of the same sex(Zekvel et al., 2014).

This finding was also supported by studies employing the ITFS procedure (Brungart et al., 2009; Kidd et al., 2016; Rennies et al., 2019). Kidd et al. (2016) observing a large masking release of 21.4 dB in the different-sex condition relative to the same-sex condition. After the stimuli were processed with ITFS, a large difference in SRTs (about 30 dB) was observed between unprocessed and processed conditions for the same-sex masker whereas little (<2 dB) was for different-sex masker. Because ITFS was assumed to remove most of the IM component, the difference observed between same- and different-sex masker conditions indicated the difference in the amount of IM; that is, same-sex masker induces a great amount of IM while different-sex masker barely does.

To date, most studies investigated effects of the difference in voice characteristics on speech-on-speech masking by varying the relative sex of target and masker. Although this manipulation maximally preserves the naturalness of human voices, it has inevitably led to changes in multiple aspects of voice characteristics at the same time (such as on F0 and vocal track length, VTL). Some well-controlled experiments have attempted to parse these effects (e.g., Darwin et al., 2003; Boghdady et al., 2019). For example, Darwin et al. (2003) measured the effects of F0 and VTL differences on segregating target and masker speech. They observed that increasing differences in F0 and in VTL provided systematic improvements in speech perception, yet neither of them could fully govern the total masking releasee due to target–masker sex differences. Başkent and Gaudrain (2016) systematically varied F0 and VTL to differentiate the target voice from the masker voice in sentence perception tasks and observed a strong benefit from musical training, that is, musicians outperformed non-musician listeners. Such advantages in the speech perception of musicians were likely due to better stream segregation or enhanced cognitive functions (Başkent and Gaudrain, 2016).

3.2. High-level unmasking factors

3.2.1. Linguistic variables

Evidence has shown that when different auditory objects (or streams) compete for attentional resources (i.e., object selection, Shinn-Cunningham, 2008), linguistic factors that activate linguistic processing can exert a strong effect. These factors examined in studies include background meaningfulness (e.g., Freyman et al., 2001; Swaminathan et al., 2015; Kidd et al., 2016; Rennies et al., 2019), language familiarity (Freyman et al., 2001; Rhebergen et al., 2005; Van Engen and Bradlow, 2007; Calandruccio et al. 2010, 2013), and semantic contexts (e.g., Cooke et al., 2008; Calandruccio et al., 2010; 2013; Brouwer et al., 2012; Kidd et al., 2014; Newman, 2009; Newman et al., 2015). Viswanathan et al. (2016) referred to the benefit provided by linguistic unmasking factors as "*linguistic release from masking*". A theory that summarizes the overall effect of these factors is the *target-masker linguistic similarity hypothesis* (Brouwer et al., 2012) in which it was assumed that the more similar the target and the masker speech are in these linguistic aspects, the harder it is to perform stream segregation effectively. This theory has been supported by findings from many aforementioned studies, although some other studies (e.g., Kidd et al., 2008) adapting the "every-other-word" paradigm found little evidence in support of this theory. In Kidd et al. (2008) study, the masking effect was only observed when the manipulation of linguistic variables was performed on the target sentences but not on the masker sentences, while the target-masker similarity remained the same.

3.2.1.1. Masker meaningfulness. Time reversal of an intelligible speech preserves its original spectral and temporal envelope properties while eliminating the meaning of speech (cf. Kellogg 1939; Cherry 1953; Schubert and Schultz 1962). Recent studies also confirmed that the amount of EM in a time-forward speech and in a time-reversed speech was about the same (cf. Marrone et al., 2008; Kidd et al., 2016). Therefore, timely reversing speech is an appropriate way to eliminate the meaningfulness of intelligible speech and allows the examination of lexicality in speech-on-speech masking.

In an early study by Freyman et al. (2001), the performance in nonsense sentence identification for a time-forward two-talker masker was compared with that for a timely-reversed version of the masker. The amount of IM was estimated by measuring SRM (i.e., the shift in dB between co-located and spatial separated conditions for equal recognition scores on the psychometric function). The intelligible two-talker masker induced 6 dB more SRM than the time-reversed masker, indicating that the background meaningfulness in speech-on-speech masking is a significant source of IM and the timely-reversed maskers reduce IM. Consistent findings were also reported by Kidd et al. (2008). They employed an "every-other word" paradigm in which the five-word matrix sentences (of the structure 'name, verb, number, adjective, object'), with target words forming odd-number elements whereas masker words or time-reversed words or noise bursts forming even-number elements, were used. The results showed that natural maskers caused the most reduction in identification performance, timely-reversed maskers worsened the performance by a smaller degree, and noise burst did not affect the performance. As there is no target/masker overlap, reduction in performance relative to no masker condition can be solely attributed to IM.

In Kidd et al. (2016), stimuli were processed using ITFS (i.e., glimpsed conditions) prior to the speech perception task. It was found that the difference in performance between the time-forward and time-reversed four-talker masker conditions was about 15 dB. As glimpsed stimuli preserve little IM component, the difference between unprocessed and glimpsed conditions served as an estimation of the amount of IM. For two-talker masker conditions, the estimated IM was about 30 dB, whereas for the timely-reversed condition it was about 12 dB. Kidd et al. (2016) explained that the IM component of the time-reversed speech masker may be associated with difficulty in stream segregation because (1) time-reversed speech contains many properties as natural speech does (thus high similarity) and (2) few simple perceptual cues (such as spatial location or F0) are available in time-reversed speech. A significant amount of release from IM by timely reversing the masker has also been observed in many other studies (Marrone et al., 2008; Best et al., 2012; Kidd et al., 2010, 2016; Swaminathan et al., 2015; Ueda et al., 2017; Rennies et al., 2019).

Collectively, these findings have confirmed that the meaningfulness of a speech masker can be a significant source of IM. The observation that a speech masker of which the meaningfulness is removed can also induce IM (although by a much smaller degree, e.g., Kidd et al., 2016) may also be worth further elucidation. The underlying reasons could be essential to our understanding of the time-reversal method and relevant speech properties that contribute to IM.

3.2.1.2. Familiarity of masker language. The effect of familiarity of masking language has been examined in many different languages and there are some discrepancies in the results across studies presumably due to the difference in methodology. The effect size of background language familiarity reported in some early studies was relatively small or non-significant (Freyman et al., 2001; Rhebergen et al., 2005; Mattys et al., 2009, 2010). Conversely, some other studies suggested a strong effect of the familiarity of masking language (Van Engen and Bradlow, 2007; Garcia- Lecumberri and Cooke, 2006; Clandruccio et al., 2010, 2013).

Van Engen and Bradlow (2007) found a significant masking release in identifying open-set meaningful English sentences for native English speakers when two-talker babble masker was spoken in Mandarin Chinese instead of English at challenging SNRs (i.e., −5 dB). Mattys et al. (2010), who observed a non-significant effect of masking language familiarity with close-set format tasks, carefully compared their design features to those of Van Engen and Bradlow (2007) and concluded that the inconsistent conclusions may result from a difference in the degree of required task-related attention between close- and open-set perception tasks. In other words, the independent effect of background language familiarity exists under the constraints of task-related attentional factors (Brouwer et al., 2012).

3.2.1.3. Semantic information. In section 3.2.1.1 we discussed the effect of masker meaningfulness on IM by focusing on studies that used the time-reversed speech to eliminate the meaning of the natural speech signal. In this section, the focus will be on studies in which the meaningfulness of natural speech is preserved to various degrees in terms of semantic information. It is shown that semantic contents of speech may strongly impact speech perception performance especially when the listening environment is not ideal (e.g., Freyman 2001; Rhebergena et al., 2005; Cooke et al., 2008; Brouwer et al., 2012; Calandruccio et al., 2010; 2013; 2017; Kidd et al., 2014; Viswanathan et al., 2016). Essentially, the degree of semantic information present in speech is associated with the predictability of an unfolding speech sequence (Hunter and Pisoni, 2018) by which IM can be affected.

Brouwer et al. (2012) manipulated the amount of semantic information in the target/masker similarity (i.e., meaningful sentences or syntactically correct but semantically anomalous sentences). The observed performance in identifying English target sentences was in a descending order using meaningful English sentence masker, anomalous English sentence masker, meaningful Dutch sentence masker, and anomalous Dutch sentence masker. This pattern was more obvious when the SNR was more challenging (i.e., −5 dB). Consistent results were also observed in a similar but separate study where the effect of EM was accounted for (Calandruccio et al., 2013). By using three maskers in different languages which differed in the typological distance to English (far: Mandarin Chinese; closer: Dutch; closest: English), the identification of English sentences in native English speakers was found to be best using Mandarin Chinese maskers and worst using English maskers, indicating that a linguistically dissimilar masker (Mandarin Chinese) has induced less IM than a linguistically similar masker (English).

In summary, linguistic variables influence the degree of IM in speech-on-speech masking conditions, however, these effects seem to be sensitive to the design and methods of experiments. The differences in methodology across studies may have contributed to the inconsistency in the findings regarding the effect of linguistic factors on speech unmasking. Although the linguistic similarity theory can govern a broad explanation of how linguistic variables affect speech perception, the extent to which each of those factors affects IM needs more detailed investigation while controlling for the variability in experimental designs.

3.2.2. *Attention and cognitive factors*

Although attention is not required in automatic sound segregation (Sussman, 2005), it can be a strong unmasking factor by reducing EM/MM, and IM. Studies using EEG (for details see Sussman, 2017; Fritz et al., 2007) demonstrated that attention facilitates sound segregation towards task-specific goals (Sussman, 2007) and helps with overcoming noisy listening environments (Sussman and Steinschneider, 2009). This is also shown in many behavioral studies using priming (e.g., Freyman et al., 2004; Best et al., 2007; Ihlefeld and Shinn-Cunningham, 2008; Jones and Litovsky, 2008; Kopco et al., 2009; Kitterick et al., 2010; Kidd et al., 2005; 2014; Singh et al., 2008; Carlile and Corkhill, 2015). Priming is a psychological experimental paradigm in which prior knowledge of some perceptual property of the target was provided. Essentially, cueing or priming provides a perceptual goal for listeners to focus their attention on.

In Freyman et al. (2004), when listeners were primed with sentences spoken by a different talker or printed and read silently, a significant improvement in their perception of nonsense sentences was observed. Kidd et al. (2005) instructed participants to listen to three different sentences chosen from CRM played concurrently over three spatially separated loudspeakers. In one experiment, the target was randomly assigned to one of the three loudspeakers, and the call sign was given after the three sentences were presented. Listeners were able to correctly identify the color and number associated with the call sign approximately 1/3 of the time. However, when prior knowledge about target location (i.e., which loudspeaker) was given, participants were able to correctly report the color and number over 90% of the time. It was concluded that providing prior knowledge about target source location can improve speech recognition in speech-on-speech masking (Kidd et al., 2005). Similarly, in other studies, cueing listeners as to ‘who’ or ‘what’ to listen for also provided significant benefit in speech perception in noise compared to situations where listeners were not provided with any clues about the target speech.

On the other hand, as the total cognitive resources or capacity is limited (Murphy et al., 2017), increasing cognitive loads by adding a secondary task that requires attentional processing (e.g., Mattys et al., 2009; Sörqvist and Rönnberg 2014; Baldock et al., 2019) or demands working memory (e.g., Francis, 2010; Brungart et al., 2013) can leave speech perception open to noise interference. In cocktail party listening environments, listeners may perform two types of attentional processing depending on their perceptual goal: selective attention (e.g., focusing on only one talker while ignoring others) or divided attention (i.e., extract information across many talkers). The latter was found to produce higher cognitive loads as measured by pupil responses (see a review by Zekveld et al., 2018). Baldock et al. (2019) had each participant undertake selective and divided auditory attention tasks using the dichotic digits test. Participants were asked to repeat numbers played between left and right ears (i.e., divided attention) or only in one ear (i.e., selective attention). Pupil responses were measured during these tasks. Larger mean and peak pupil dilation (indicative of greater cognitive load) were observed when listeners were asked to divide their attention across two ears than in one ear. This finding suggested that limited cognitive capacity or cognitive deficits can result in poor performance in speech perception tasks involving divided attention.

Brungart et al. (2013) measured speech recognition performance in different listening conditions (target in one known ear, in one unknown ear, or in both ears, masked by SSN or speech babbles). The recognition threshold generally worsened from response to one known ear to both ears for all maskers, but the decrement was much larger for speech babbles than SSN. In the second experiment, listeners were asked to perform a secondary memory task (i.e., one-back task) in which they were required to remember the last word of a sentence, while doing a true-false judgment of the sentence in the presence of speech babbles or noise. The observed performance for speech babble masker was significantly worse than that with noise. Brungart et al. (2103) concluded that speech-on-speech masking tasks require more cognitive resources than speech-in-noise tasks as additional cognitive resources need to be allocated to deal with IM in speech babbles. In line with this finding, other studies also reported worse speech recognition performance associated with higher cognitive load, likely due to that attention- or working-memory-demanding tasks preempt processing resources needed for speech perception (Mattys and Wiget, 2011; Mattys and Scharenborg, 2014; Mattys et al., 2013).

These studies showed that although attention can guide low-level processing in speech perception (such as stream segregation) towards the listener’s perceptual goal or improve its efficiency, it also draws upon limited cognitive resources, especially when the task involves divided attention. Extracting speech information from other speech seems to be more demanding on cognitive resources than from noise. Besides, when listeners need to perform multiple tasks that involve different levels of processing, the distribution of cognitive recourse seems to prioritize those requiring higher-level processing.

Overall, English research has revealed that spatial separation and difference in voice characteristics between target and masker in cocktail party environments can facilitate speech perception by a relatively large degree by reducing IM at early stages of auditory processing (pre-attentional). Linguistic factors also play important roles in speech unmasking depending on the degree of linguistic similarity between the target and masker. The influence of these factors can be modulated or overwhelmed by attention whereas priming can help overcome multi-talker distraction. Overall, speech processing is limited by cognitive resources. Listeners tend to prioritize attentional processing over other lower-level processing.

4. Speech unmasking in Mandarin Chinese

Most studies investigating masking and unmasking factors in cocktail party problems have been conducted in English or other non-tonal languages. Tonal languages, as a major part of world languages, have been less frequently used for the investigation of this topic. As one of those tonal languages, Mandarin Chinese has four lexical tones that function similarly to vowels and consonants in English, that is, to discriminate the lexical meaning of monosyllables (Liang, 1963; Xu and Zhou, 2011). This section reviews findings on unmasking factors in Mandarin Chinese.

4.1. Spatial separation

Wu et al. (2005) replicated the experiment by Freyman et al. (1999) in order to investigate the amount of SRM that existed in Mandarin Chinese. For the two-talker masker, SRM for Mandarin Chinese nonsense sentences was about 3.3 dB which was much smaller than 9 dB that was observed by Freyman et al. (1999) and many other recent studies in English.

In Wu et al. (2007), the investigation of SRM was extended to multiple-talker maskers (number of talkers (N): 1, 2, 3, or 4) for Mandarin Chinese to compare to Freyman et al. (2004). Results showed that SRM at N = 1 for Mandarin Chinese was very small, however, in a similar condition, Freyman et al. (2004) observed a 7-dB SRM. The SRM at N = 2, 3, and 4 ranged from 3.86 to 2.62 dB in Wu et al. (2007) compared to 9 to 4 dB in Freyman et al. (2004), respectively. Again, SRM for Mandarin Chinese was much smaller than that for English. Wu and colleagues speculated that the relatively smaller benefits of perceived spatial separation on reducing IM in Mandarin Chinese may be due to that Chinese syllables have more voiceless consonants and fewer voiced consonants than English syllables, thus, the Chinese language has a larger energetic masking component (Kang, 1998). Mandarin Chinese speech may have a different target/masker similarity pattern than English speech. Lastly, the pitch contour of a vowel is phonemic in Mandarin Chinese but not in English.

In a recent study by Zhang et al. (2020), the magnitude of SRM (measured in SRTs) between Mandarin-Chinese-speaking and English-speaking listeners was compared using close-set matrix-style test materials as stimuli in corresponding languages. Contrary to Wu et al. (2005, 2007), comparable magnitudes in SRM between Mandarin Chinese and English speakers were observed. The Chinese listeners benefited by about 13.7 dB and comparably, English listeners by about 12.2 dB from the spatial separation alone. It should be noted that there were several methodological differences between Zhang et al. (2020) and Wu et al. (2005, 2007). Zhang et al. (2020) manipulated the physical location of loudspeakers, whereas in Wu et al. (2005, 2007), the precedence effect was used to induce the spatial separation. Also, Zhang et al. (2020) used matrix sentences as the test material, while nonsense sentences were used in Wu et al. (2005, 2007).

4.2. Speaker voice difference

Chen et al. (2020) measured SRTs with one, two, or four masker talkers for different combinations of target-masker sexes. A masking release of about 12 dB was observed in the different-sex masker condition compared to the same-sex condition. Zhang et al. (2020) reported a masking release of 11.7 dB when talker sex cues were available in Chinese listeners, which was significantly better than that of English listeners in their study (about 8.7 dB). The authors attributed the larger masking release in Mandarin Chinese listeners to their tonal language experience which might have facilitated accurate talker identification and pitch perception (Xie and Myers, 2015; Deroche et al., 2019).

4.3. Linguistic variables

Wu et al. (2011) compared the word recognition for different target/masker language combinations in native Chinese and English speakers: same-language (e.g., target and masker both in Mandarin Chinese or English), cross-language (e.g., target in English masker in Mandarin Chinese, or reversely) and noise conditions (i.e., SSN as masker). Results showed that Chinese listeners benefited from spatial separation as much as English listeners did for the SSN masker, but they benefited less in their same-language condition (i.e., Chinese target and Chinese masker). Interestingly, when Chinese target words were scored as correct when either of the two morphemes in the target word was correctly identified (i.e., morphemic-level scoring) instead of whole-word scoring, the reduced benefit from spatial separation in Mandarin Chinese listeners was comparable. Note that the morphemic-level scoring did not change the psychometric function for English listeners' identification scores. Wu and colleagues concluded that release from IM happens at the morphemic level for Mandarin Chinese and spatial separation may facilitate morpheme access. They also suspected that the reason underlying the difference in SRM between the two languages using whole-word scoring was that most Chinese multisyllabic words are also multimorphemic (meaning that each monosyllable can also stand alone as a new word). On the contrary, most English multisyllabic words are monomorphemic, which means the lexical meaning of those words is carried by several syllables together and each syllable cannot stand alone as a new word. This difference, according to Wu et al. (2011) may have led to different ways in which lexical meaning was accessed in the two languages.

Peng et al. (2012) compared word and sentence recognition in Mandarin Chinese with multi-talker maskers. Sentence recognition scores were significantly higher than those of word recognition. The authors suggested that the linguistic connection in sentences might have helped listeners perceiving the target better in background babbles.

4.4. F0 contours

In Mandarin Chinese, F0 contour is the primary cue for tone perception at the monosyllable level (Howie, 1976; Xu and Zhou, 2011). It is also related to the perception of sentence intonation and voice pitch. Many studies have focused their interest on the contribution of F0 contours in perceiving Mandarin Chinese (e.g., Kong and Zeng, 2006; Krenmayr et al., 2011; Lee et al., 2013; Li et al., 2019; Wang and Xu, 2020). Using meaningful sentences as targets, Patel et al. (2010) demonstrated that meaningful Chinese sentences with flattened F0 contours were as intelligible as those with normal F0 patterns in a quiet environment, however, it was less intelligible under SSN or babble noise, as reported in several other studies (e.g., Wang et al., 2013; Chen et al., 2014; Li et al., 2019).

To control for other linguistic effects and to investigate whether intonation cues affect the unmasking of Mandarin Chinese speech, Wu (2019) manipulated the intonation information of nonsense Mandarin Chinese sentences in three conditions (i.e., flattened, typical, and exaggerated intonation) while preserving the tone information of each monosyllable. The recognition performance was measured in the presence of steady-state noise or two-talker babble. Consistent with English findings (Laures, 1999; Laures and Bunton, 2003; Binns and Culling, 2007), natural intonation information provided the largest benefit in speech perception in noise. Furthermore, Wu's (2019) results suggested that either reducing or exaggerating intonation relative to natural intonation reduced speech intelligibility, especially for the multi-talker babble condition.

4.5. Target priming

In Yang et al. (2007), the effect of voice cueing on releasing Chinese nonsense sentences from IM was examined by employing a same-sentence, different-sentence priming, and no priming condition. All primers were spoken by the same voice as in the target. Yang and colleagues found that when the masker was speech, even though the content of primer was not the same as the target, listeners could still benefit from the same voice cue (i.e., different-sentence condition). They concluded that similar to English, being familiar with the target voice also benefited Mandarin Chinese speech perception in speech-on-speech masking. Voice cues, which act at the perceptual level, can facilitate selective attention to the voice characteristics of the target stream, leading to a release from IM. To account for any potential effect due to long-term familiarity of the target voice, Huang et al. (2010) used both single and double presentations of priming sentences to investigate how listeners utilized voice information to reduce IM. Consistent with Yang et al. (2007), a significant benefit of being familiar with the target voice was observed in the perception of Mandarin Chinese nonsense sentences for adult listeners.

The effect of content-priming (i.e., presenting the early part of a target sentence in quiet) was also observed in perceiving nonsense Mandarin Chinese sentences in noise (Wu et al., 2012a, 2012b). Wu et al. (2012a) observed that participants benefited from the content priming by about 2 dB in the presence of a multi-talker masker in reporting the last target keyword. The authors agreed with the English literature (Freyman et al., 2004) that the content prime mainly helps listeners focus attention more quickly on the target, thereby facilitating recognition of the last keyword in the target stream against IM.

In summary, Mandarin Chinese, as a tonal language, can also benefit from factors that have been shown to unmask speech perception in cocktail-party listening environments in English. However, this line of research was relatively sparse. The reported effects of low-level unmasking factors such as spatial separation and difference in voice characteristics were inconsistent across the few studies and some of them appeared to lack replicability. For high-level unmasking factors, findings in the few Mandarin Chinese reports seemed to be consistent with those in English. Yet, most studies focused on replicating English studies. Systematic investigations for attentional factors were absent. In addition, as a unique feature for tonal language, the pitch contour that plays multiple roles in Mandarin Chinese was found to be associated with speech unmasking, however, more detailed research is needed to expand our knowledge on this matter.

5. Summary

Auditory masking related to speech perception in cocktail-party listening conditions can be classified in several forms. In particular, energetic masking and modulation masking are characterized by peripheral interference. Informational masking is characterized by the "confusion" that occurs beyond the auditory peripheral level.

Both English and Mandarin Chinese literature has demonstrated that speech perception in cocktail-party listening environments can be unmasked by several factors. Some of the factors facilitate low-level auditory processing (e.g., spatial separation and voice characteristics) and some other factors benefit high-level processing (e.g., linguistic variables, attention, and cognitive capacity). Compared to English research, fewer studies have been done on speech unmasking in Mandarin Chinese and little consensus regarding the magnitude of unmasking benefits due to spatial separation and voice characteristics has been reached.

Interestingly, some studies have suggested that linguistic differences between Mandarin Chinese and English may result in differences in the degree to which two languages benefit from low-level unmasking factors. As Mandarin Chinese has fewer voiced consonants than English, it may be more subject to energetic masking than English (Wu et al., 2005, 2011), leading to different distributions of energetic masking or modulation masking and informational masking. This in turn may be reflected in the magnitude of the spatial release of masking. For voice characteristics, since pitch contours play multiple roles in tonal language, the observed masking release due to voice difference may differ from that found in English. Indeed, early studies in Mandarin Chinese have shown such possibilities. Further investigation is encouraged in light of the linguistic differences between tonal and non-tonal languages.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

None.

Acknowledgments

We thank Devon Zurovchak for her assistance in proofreading the manuscript.

References

Allen, K., Carlile, S., Alais, D., 2008. Contributions of talker characteristics and spatial location to auditory streaming. J. Acoust. Soc. Am. 123, 1562–1570. https://doi.org/10.1121/1.2831774.

Andéol, G., Suied, C., Scannella, S., Dehais, F., 2017. The spatial release of cognitive load in cocktail party is determined by the relative levels of the talkers. J. Assoc. Res. Otolaryngol. 18 (3), 457–464. https://doi.org/10.1007/s10162-016-0611-7.

Arbogast, T.L., Mason, C.R., Kidd Jr., G., 2002. The effect of spatial separation on informational and energetic masking of speech. J. Acoust. Soc. Am. 112 (5), 2086–2098. https://doi.org/10.1121/1.1510141.

Baldock, J., Kapadia, S., van Steenbrugge, W., 2019. The task-evoked pupil response in divided auditory attention tasks. J. Am. Acad. Audiol. 30 (4), 264–272. https://doi.org/10.3766/jaaa.17060.

Baskent, D., Gaudrain, E., 2016. Musician advantage for speech-on- speech perception. J. Acoust. Soc. Am. 139, EL51–EL56. https://doi.org/10.1121/1.4942628.

Best, V., Gallun, F.J., Ihlefeld, A., Shinn-Cunningham, B.G., 2006. The influence of spatial separation on divided listening. J. Acoust. Soc. Am. 120, 1506–1516. https://doi.org/10.1121/1.2234849.

Best, V., Marrone, N., Mason, C.R., Kidd Jr., G., 2012. The influence of non-spatial factors on measures of spatial release from masking. J. Acoust. Soc. Am. 131 (4), 3103–3110. https://doi.org/10.1121/1.3693656.

Best, V., Ozmeral, E.J., Shinn-Cunningham, B.G., 2007. Visually-guided attention enhances target identification in a complex auditory scene. J. Assoc. Res. Otolaryngol. 8 (2), 294–304. https://doi.org/10.1007/s10162-007-0073-z.

Binns, C., Culling, J.F., 2007. The role of fundamental frequency contours in the perception of speech against interfering speech. J. Acoust. Soc. Am. 122 (3), 1765–1776. https://doi.org/10.1121/1.2751394.

Boghdady, N.E., Gaudrain, E., Başkent, D., 2019. Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users? J. Acoust. Soc. Am. 145 (1), 417–439. https://doi.org/10.1121/1.5087693.

Bronkhorst, A., 2000. The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. Acta Acust. United Ac. 86, 117–128. https://doi.org/10.3758/s13414-015-0882-9.

Bronkhorst, A., 2015. The cocktail-party problem revisited: early processing and selection of multi-talker speech. Atten. Percept. Psychophys. 77 (5), 1465–1487. https://doi.org/10.3758/s13414-015-0882-9.

Brouwer, S., Van Engen, K., Calandruccio, L., Bradlow, A.R., 2012. Linguistic contributions to speech-on-speech masking for native and non-native listeners: language familiarity and semantic content. J. Acoust. Soc. Am. 131 (2), 1449–1464. https://doi.org/10.1121/1.3675943.

Brown, D.K., Cameron, S., Martin, J., Watson, C., Dillon, H., 2010. The North American Listening in Spatialized Noise-Sentences Test (NA LiSN-S): normative data and test-retest reliability studies for adolescents and young adults. J. Am. Acad. Audiol. 21 (10), 629–641. https://doi.org/10.3766/jaaa.21.10.3.

Brown, G.J., Cooke, M., 1994. Computational auditory scene analysis. Comput. Speech Lang 8, 297–336. https://doi.org/10.1006/csla.1994.1016.

Brungart, D.S., 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Am. 109 (3), 1101–1109. https://doi.org/10.1121/1.1345696.

Brungart, D.S., Simpson, B.D., 2002. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. J. Acoust. Soc. Am. 112, 664–676. https://doi.org/10.1121/1.1490592.

Brungart, D.S., Chang, P.S., Simpson, B.D., Wang, D., 2006. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Am. 120 (6), 4007–4018. https://doi.org/10.1121/1.2363929.

Brungart, D.S., Chang, P.S., Simpson, B.D., Wang, D., 2009. Multitalker speech perception with ideal time-frequency segregation: effects of voice characteristics and number of talkers. J. Acoust. Soc. Am. 125 (6), 4006–4022. https://doi.org/10.1121/1.3117686.

Brungart, D.S., Simpson, B.D., Ericson, M.A., Scott, K.R., 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. J. Acoust. Soc. Am. 110 (5), 2527–2538. https://doi.org/10.1121/1.1345696.

Brungart, D., Iyer, N., Thompson, E., Simpson, B.D., Gordon-Salant, S., Shurman, J., Grant, K.W., 2013. Interactions between listening effort and masker type on the energetic and informational masking of speech stimuli. J. Acoust. Soc. Am. 133 (5) https://doi.org/10.1121/1.4806059, 3435-3435.

Calandruccio, L., Brouwer, S., Van Engen, K., Dhar, S., Bradlow, A., 2013. Masking release due to linguistic and phonetic dissimilarity between the target and masker speech. Am. J. Audiol. 22 (1), 157–164. https://doi.org/10.1044/1059-0889(2013/12-0072.

Calandruccio, L., Buss, E., Bowdrie, K., 2017. Effectiveness of two-talker maskers that differ in talker congruity and perceptual similarity to the target speech. Trends Hear 21. https://doi.org/10.1177/2331216517709385.

Calandruccio, L., Dhar, S., Bradlow, A.R., 2010. Speech-on-speech masking with variable access to the linguistic content of the masker speech. J. Acoust. Soc. Am. 128 (2), 860–869. https://doi.org/10.1121/1.3458857.

Carhart, R., Tillman, T.W., Greetis, E.S., 1968. Perceptual masking in multiple sound backgrounds. J. Acoust. Soc. Am. 45, 694–703. https://doi.org/10.1121/1.1911445.

Carlile, S., Corkhill, C., 2015. Selective spatial attention modulates bottom-up informational masking of speech. Sci. Rep. 5, 8662. https://doi.org/10.1038/srep08662.

Chen, B., Shi, Y., Zhang, L., Sun, Z., Li, Y., Gopen, Q., Fu, Q.J., 2020. Masking effects in the perception of multiple simultaneous talkers in normal-hearing and cochlear implant listeners. Trends Hear 24. https://doi.org/10.1177/2331216520916106, 2331216520916106.

Chen, F., Wong, L., Hu, Y., 2014. Effects of lexical tone contour on Mandarin sentence intelligibility. J. Speech Lang. Hear. Res. 57 (1), 338–345. https://doi.org/10.1044/1092-4388(2013/12-0324.

Cherry, C.E., 1953. Some experiments on the recognition of speech, with one and two ears. J. Acoust. Soc. Am. 25, 975–979. https://doi.org/10.1121/1.1907229.

Cooke, M., Lecumberri, M.G., Barker, J., 2008. The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. J. Acoust. Soc. Am. 123 (1), 414–427. https://doi.org/10.1121/1.2804952.

Culling, J.F., Stone, M.A., 2017. Energetic masking and masking release. In: Middlebrooks, J.C., Simon, J.Z., Popper, A.N., Fay, R.R. (Eds.), The Auditory System at the Cocktail Party. Springer Handbook of Auditory Research. Springer International Publishing, New York, NY, pp. 1–6. https://doi.org/10.1007/978-3-319-51662-2_1.

Culling, J.F., Hawley, M.L., Litovsky, R.Y., 2004. The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. J. Acoust. Soc. Am. 116 (2), 1057–1065. https://doi.org/10.1121/1.1772396.

Culling, J.F., Mansell, E.R., 2013. Speech intelligibility among modulated and spatially distributed noise sources. J. Acoust. Soc. Am. 133 (4), 2254–2261. https://doi.org/10.1121/1.4794384.

Cullington, H.E., Zeng, F., 2008. Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects. J. Acoust. Soc. Am. 123 (1), 450–461. https://doi.org/10.1121/1.2805617.

Darwin, C.J., Brungart, D.S., Simpson, B.D., 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. J. Acoust. Soc. Am. 114, 2913–2922. https://doi.org/10.1121/1.1616924.

Deroche, M.L.D., Lu, H.P., Kulkarni, A.M., Caldwell, M., Barrett, K.C., Peng, S.C., Limb, C.J., Lin, Y.S., Chatterjee, M., 2019. A tonal-language benefit for pitch in normally-hearing and cochlear-implanted children. Sci. Rep. 9, 109. https://doi.org/10.1038/s41598-018-36393-1.

Diehl, R.L., 2008. Acoustic and auditory phonetics: the adaptive design of speech sound systems. Philos. Trans. R. Soc. B. 363, 965–978. https://doi.org/10.1098/rstb.2007.2153.

Durlach, N., 2006. Auditory masking: need for improved conceptual structure. J. Acoust. Soc. Am. 120, 1787–1790. https://doi.org/10.1121/1.2335426.

Edmonds, B.A., Culling, J.F., 2006. The spatial unmasking of speech: evidence for better-ear listening. J. Acoust. Soc. Am. 120 (3), 1539–1545. https://doi.org/10.1121/1.2228573.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton, The Hague, The Netherlands, pp. 15–90.

Fletcher, H., 1940. Auditory patterns. Rev. Mod. Phys. 12 (1), 47–65. https://doi.org/10.1103/RevModPhys.12.47.

Francis, A.L., 2010. Improved segregation of simultaneous talkers differentially affects perceptual and cognitive capacity demands for recognizing speech in competing speech. Atten. Percept. Psychophys. 72, 501–516. https://doi.org/10.3758/APP.72.2.501.

Freyman, R.L., Balakrishnan, U., Helfer, K.S., 2001. Spatial release from informational masking in speech recognition. J. Acoust. Soc. Am. 109 (5), 2112–2122. https://doi.org/10.1121/1.1354984.

Freyman, R.L., Balakrishnan, U., Helfer, K.S., 2004. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. J. Acoust. Soc. Am. 115 (5), 2246–2256. https://doi.org/10.1121/1.1689343.

Freyman, R.L., Helfer, K.S., McCall, D.D., Clifton, R.K., 1999. The role of perceived spatial separation in the unmasking of speech. J. Acoust. Soc. Am. 106 (6), 3578–3588. https://doi.org/10.1121/1.428211.

Fritz, J.B., Elhilali, M., David, S.V., Shamma, S.A., 2007. Auditory attention: focusing the searchlight on sound. Curr. Opin. Neurol. 17 (4), 437–455. https://doi.org/10.1016/j.conb.2007.07.011.

Gallun, F.J., Kampel, S.D., Diedesch, A.C., Jakien, K.M., 2013. Independent impacts of age and hearing loss on spatial release in a complex auditory environment. Front. Neurosci. 252, 1–11. https://doi.org/10.3389/fnins.2013.00252.

Garcia Lecumberri, M.L., Cooke, M., 2006. Effect of masker type on native and non-native consonant perception in noise. J. Acoust. Soc. Am. 119 (4), 2445–2454. https://doi.org/10.1121/1.2180210.

Hirsh, I.J., 1948. The influence of interaural phase on interaural summation and inhibition. J. Acoust. Soc. Am. 20, 536–544. https://doi.org/10.1121/1.1906407.

Hoen, M., Meunier, F., Grataloup, C.-L., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X., Collet, L., 2007. Phonetic and lexical interferences in informational masking during speech-in-speech comprehension. Speech Commun. 49, 905–916. https://doi.org/10.1016/j.specom.2007.05.008.

Howie, J.M., 1976. Acoustical Studies of Mandarin Vowels and Tones. Cambridge University Press, Cambridge [England]. https://doi.org/10.2307/600754.

Huang, Y., Xu, L., Wu, X., Li, L., 2010. The effect of voice cuing on releasing speech from informational masking disappears in older adults. Ear Hear. 31, 579–583. https://doi.org/10.1097/AUD.0b013e3181db6dc2.

Hunter, C.R., Pisoni, D.B., 2018. Extrinsic cognitive load impairs spoken word recognition in high- and low-predictability sentences. Ear Hear. 39 (2), 378–389. https://doi.org/10.1097/AUD.0000000000000493.

Ihlefeld, A., Shinn-Cunningham, B., 2008. Disentangling the effects of spatial cues on selection and formation of auditory objects. J. Acoust. Soc. Am. 124 (4), 2224–2235. https://doi.org/10.1121/1.2973185.

Jones, G.L., Litovsky, R.Y., 2008. Effects of uncertainty in a cocktail party environment in adults. J. Acoust. Soc. Am. 124, 3818–3830.

Kang, J., 1998. Comparison of speech intelligibility between English and Chinese. J. Acoust. Soc. Am. 103 (2), 1213–1216. https://doi.org/10.1121/1.421253.

Kellogg, E.W., 1939. Reversed speech. J. Acoust. Soc. Am. 10 (4), 324–326.

Kidd Jr., G., Arbogast, T.L., Mason, C.R., Gallun, F.J., 2005. The advantage of knowing where to listen. J. Acoust. Soc. Am. 118 (6), 3804–3815. https://doi.org/10.1121/1.2109187.

Kidd Jr., G., Best, V., Mason, C.R., 2008. Listening to every other word: examining the strength of linkage variables in forming streams of speech. J. Acoust. Soc. Am. 124 (6), 3793–3802. https://doi.org/10.1121/1.2998980.

Kidd Jr., G., Mason, C.R., Best, V., 2014. The role of syntax in maintaining the integrity of streams of speech. J. Acoust. Soc. Am. 135 (2), 766–777. https://doi.org/10.1121/1.4861354.

Kidd Jr., G., Mason, C.R., Best, V., Marrone, N., 2010. Stimulus factors influencing spatial release from speech-on-speech masking. J. Acoust. Soc. Am. 128 (4), 1965–1978. https://doi.org/10.1121/1.3478781.

Kidd, G., Mason, C.R., Best, V., Roverud, E., Swaminathan, J., Jennings, T., 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 (1), 440–457. https://doi.org/10.1121/1.5087555.

Kidd Jr., G., Mason, C.R., Swaminathan, J., Roverud, E., et al., 2016. Determining the energetic and informational components of speech-on-speech masking. J. Acoust. Soc. Am. 140 (1), 132–144. https://doi.org/10.1121/1.4954748.

Kitterick, P.T., Bailey, P.J., Summerfield, A.Q., 2010. Benefits of knowing who, where, and when in multi-talker listening. J. Acoust. Soc. Am. 127 (4), 2498–2508. https://doi.org/10.1121/1.3327507.

Kong, Y.Y., Zeng, F.G., 2006. Temporal and spectral cues in Mandarin tone recognition. J. Acoust. Soc. Am. 120 (5), 2830–2840. https://doi.org/10.1121/1.2346009.

Kopčo, N., Best, V., Carlile, S., 2009. Localizing a speech target in a multitalker mixture. J. Acoust. Soc. Am. 125 (4) https://doi.org/10.1121/1.4784289, 2691-2691.

Krenmayr, A., Qi, B., Liu, B., Liu, H., Chen, X., Han, D., Zierhofer, C.M., 2011. Development of a Mandarin tone identification test: sensitivity index d' as a performance measure for individual tones. Int. J. Audiol. 50 (3), 155–163. https://doi.org/10.3109/14992027.2010.530613.

Laures, J.S., Bunton, K., 2003. Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. J. Commun. Disord. 36 (6), 449–464. https://doi.org/10.1016/S0021-9924(03)00032-7.

Laures, J.S., Weismer, G., 1999. The effects of a flattened fundamental frequency on intelligibility at the sentence level. J. Speech Lang. Hear. Res. 42 (5), 1148–1156. https://doi.org/10.1044/jslhr.4205.1148.

Lee, C.-Y., Tao, L., Bond, Z.S., 2013. Effects of speaker variability and noise on Mandarin tone identification by native and non-native listeners. Speech Lang. Hear. 16 (1), 46–54. https://doi.org/10.1179/2050571X12Z.0000000003.

Li, N., Wang, S., Wang, X., Xu, L., 2019. Contributions of lexical tone to Mandarin sentence recognition in hearing-impaired listeners under noisy conditions. J. Acoust. Soc. Am. 146 (2), EL99–EL105. https://doi.org/10.1179/2050571X12Z.0000000003.

Liang, Z.A., 1963. The auditory perception of Mandarin tones. Acta Physiol. Sin. 26 (2), 85–91.

Marrone, N.L., Mason, C.R., Kidd Jr., G., 2008. Tuning in the spatial dimension: evidence from a masked speech identification task. J. Acoust. Soc. Am. 124 (2), 1146–1158. https://doi.org/10.1121/1.2945710.

Mattys, S.L., Barden, K., Samuel, A.G., 2013. Extrinsic cognitive load impairs low-level speech perception. Psychon. Bull. Rev. 21 (3), 748–754. https://doi.org/10.3758/s13423-013-0544-7.

Mattys, S.L., Brooks, J., Cooke, M., 2009. Recognizing speech under a processing load: dissociating energetic from informational factors. Cognit. Psychol. 59 (3), 203–243. https://doi.org/10.1016/j.cogpsych.2009.04.001.

Mattys, S.L., Carroll, L.M., Li, C.K.W., Chan, S.L.Y., 2010. Effects of energetic and informational masking on speech segmentation by native and non-native speakers. Speech Commun. 11, 887–899. https://doi.org/10.1016/j.specom.2010.01.005.

Mattys, S., Davis, M., Bradlow, A., Scott, S., 2012. Speech recognition in adverse conditions: a review. Lang. Cognit. Process. 27 (7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006.

Mattys, S., Scharenborg, O., 2014. Phoneme categorization and discrimination in younger and older adults: a comparative analysis of perceptual, lexical, and attentional factors. Psychol. Aging 29 1, 150–162. https://doi.org/10.1037/a0035387.

Mattys, S.L., Wiget, L., 2011. Effects of cognitive load on speech recognition. J. Mem. Lang. 65 (2), 145–160. https://doi.org/10.1016/j.jml.2011.04.004.

Moore, B.C.J., Gockel, H.E., 2012. Properties of auditory stream formation. Philos. Trans. R. Soc. B. 367, 919–931. https://doi.org/10.1098/rstb.2011.0355.

Murphy, S., Spence, C., Dalton, P., 2017. Auditory perceptual load: a review. Hear. Res. 352, 40–48. https://doi.org/10.1016/j.heares.2017.02.005.

Newman, R., 2009. Infants' listening in multitalker environments: effect of the number of background talkers. Atten. Percept. Psychophys. 71 (4), 822–836. https://doi.org/10.3758/APP.71.4.822.

Newman, R.S., Morini, G., Ahsan, F., Kidd Jr., G., 2015. Linguistically-based informational masking in preschool children. J. Acoust. Soc. Am. 138 (1), EL93–EL98. https://doi.org/10.1121/1.4921677.

Noble, W., Perrett, S., 2002. Hearing speech against spatially separate competing speech versus competing noise. Percept. Psychophys. 64, 1325–1336. https://doi.org/10.3758/BF03194775.

Patel, A.D., Xu, Y., Wang, B., 2010. The role of F0 variation in the intelligibility of Mandarin sentences. In: International Conference on Speech Prosody, p. 2010. Chicago, 10.1.1.640.6758.

Peng, J., Zhang, H., Wang, Z., 2012. Chinese speech identification in multi-talker babble with diotic and dichotic listening. Sci. Bull. 57 (20), 2548–2553. https://doi.org/10.1007/s11434-012-5273-1.

Rennies, J., Best, V., Roverud, E., Kidd Jr., G., 2019. Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort. Trends Hear 23. https://doi.org/10.1177/2331216519854597, 2331216519854597.

Rhebergen, K.S., Versfeld, N.J., Dreschler, W.A., 2005. Release from informational masking by time reversal of native and non-native interfering speech. J. Acoust. Soc. Am. 118, 1274–1277. https://doi.org/10.1121/1.2000751.

Schneider, B., Li, L., Daneman, M., 2007. How competing speech interferes with speech comprehension in everyday listening situations. J. Am. Acad. Audiol. 18 (7), 559–572. https://doi.org/10.3766/jaaa.18.7.4.

Schubert, E.D., Schultz, M.C., 1962. Some aspects of binaural signal selection. J. Acoust. Soc. Am. 34 (6), 844–849. https://doi.org/10.1121/1.1918203.

Shinn-Cunningham, B.G., 2008. Object-based auditory and visual attention. Trends Cognit. Sci. 12 (5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003.

Shinn-Cunningham, B., Best, V., Lee, A.K.C., 2017. Auditory object formation and selection. In: Middlebrooks, J.C., Simon, J.Z., Popper, A.N., Fay, R.R. (Eds.), The Auditory System at the Cocktail Party. Springer Handbook of Auditory Research. Springer International Publishing, New York, NY, pp. 7–40. https://doi.org/10.1007/978-3-319-51662-2_2.

Singh, G., Pichora-Fuller, M.K., Schneider, B.A., 2008. The effect of age on auditory spatial attention in conditions of real and simulated spatial separation. J. Acoust. Soc. Am. 124 (2), 1294–1305. https://doi.org/10.1121/1.2949399.

Sörqvist, P., Rönnberg, J., 2014. Individual differences in distractibility: an update and a model. PsyCh J. 3 (1), 42–57. https://doi.org/10.1002/pchj.47.

Stone, M.A., Canavan, S., 2016. The near non-existence of "pure" energetic masking release for speech: extension to spectro-temporal modulation and glimpsing. J. Acoust. Soc. Am. 140 (2), 832–842. https://doi.org/10.1121/1.4960483.

Stone, M.A., Fullgrabe, C., Moore, B.C.J., 2012. Notionally steady background noise acts primarily as a modulation masker of speech. J. Acoust. Soc. Am. 132 (1), 317–326. https://doi.org/10.1121/1.4725766.

Sussman, E.S., 2007. A new view on the MMN and attention debate: the role of context in processing auditory events. J. Psychophysiol. 21 (3), 164–175. https://doi.org/10.1027/0269-8803.21.34.164.

Sussman, E.S., 2005. Integration and segregation in auditory scene analysis. J. Acoust. Soc. Am. 117, 1285–1298. https://doi.org/10.1121/1.1854312.

Sussman, E.S., 2017. Auditory scene analysis: an attention perspective. J. Speech Lang. Hear. Res. 60 (10), 2989–3000. https://doi.org/10.1044/2017_JSLHR-H-17-0041.

Sussman, E., Steinschneider, M., 2009. Attention effects on auditory scene analysis in children. Neuropsychology 47, 771–785. https://doi.org/10.1016/j.neuropsychologia.2008.12.007.

Swaminathan, J., Mason, C.R., Streeter, T.M., Best, V.A., Kidd Jr., G., Pate, A.D., 2015. Musical training, individual differences and the cocktail party problem. Sci. Rep. 26 (5), 1–10. https://doi.org/10.1038/srep11628.

Ueda, K., Nakajima, Y., Ellermeier, W., Kattner, F., 2017. Intelligibility of locally time-reversed speech: a multilingual comparison. Sci. Rep. 7 (1) https://doi.org/10.1038/s41598-017-01831-z.

Van Engen, K.J., Bradlow, A.R., 2007. Sentence recognition in native- and foreign-language multi-talker background noise. J. Acoust. Soc. Am. 121 (1), 519–526. https://doi.org/10.1121/1.2400666.

Viswanathan, N., Kokkinakis, K., Williams, B.T., 2016. Spatially separating language masker from target results in spatial and linguistic masking release. J. Acoust. Soc. Am. 140 (6), EL465–EL470. https://doi.org/10.1121/1.4968034.

Wang, D.L., Brown, G.J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Network. 10, 684–697. https://doi.org/10.1109/72.761727.

Wang, J., Shu, H., Zhang, L., Liu, Z., Zhang, Y., 2013. The roles of fundamental frequency contours and sentence context in Mandarin Chinese speech intelligibility. J. Acoust. Soc. Am. 134 (1), EL91–EL97. https://doi.org/10.1121/1.4811159.

Wang, X., Xu, L., 2020. Mandarin tone perception in multiple-talker babbles and speech-shaped noise. J. Acoust. Soc. Am. 147 (4), EL307–EL313. https://doi.org/10.1121/10.0001002.

Wu, M., 2019. Effect of F0 contour on perception of Mandarin Chinese speech against masking. PloS One 14 (1), e0209976. https://doi.org/10.1371/journal.pone.0209976.

Wu, M., Li, H., Hong, Z., Xian, X., Li, J., Wu, X., Li, L., 2012a. Effects of aging on the ability to benefit from prior knowledge of message content in masked speech recognition. Speech Commun. 54 (4), 529–542. https://doi.org/10.1016/j.specom.2011.11.003.

Wu, M., Li, H., Gao, Y., Lei, M., Teng, X., Wu, X., Li, L., 2012b. Adding irrelevant information to the content prime reduces the prime-induced unmasking effect on speech recognition. Hear. Res. 283 (1–2), 136–143. https://doi.org/10.1016/j.heares.2011.11.001.

Wu, X., Chen, J., Yang, Z., Huang, Q., Wang, M., Li, L., 2007. Effect of number of masking talkers on masking of Chinese speech. In: Proceedings of the Annual Conference of the International Speech Communication Association, vol. 2007. Interspeech, pp. 390–393.

Wu, X., Wang, C., Chen, J., Qu, H., Li, W., Wu, Y., Li, L., 2005. The effect of perceived spatial separation on informational masking of Chinese speech. Hear. Res. 199 (1–2), 1–10. https://doi.org/10.1016/j.heares.2004.03.010.

Wu, X., Yang, Z., Huang, Y., Chen, J., Li, L., Daneman, M., Schneider, B.A., 2011. Cross-language differences in informational masking of speech by speech: English versus Mandarin Chinese. J. Speech Lang. Hear. Res. 54 (6), 1506–1524. https://doi.org/10.1044/1092-4388(2011/10-0282.

Xia, J., Noorale, N., Kalluri, S., Edwards, B., 2015. Spatial release of cognitive load measured in a dual-task paradigm in normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 137 (4), 1888–1898. https://doi.org/10.1121/1.4916599.

Xie, X., Myers, E., 2015. The impact of musical training and tone language experience on talker identification. J. Acoust. Soc. Am. 137 (1), 419–432. https://doi.org/10.1121/1.4904699.

Xu, L., 2016. Temporal envelopes in sine-wave speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, vol. 2016. Interspeech, San Francisco, CA, pp. 1682–1686. https://doi.org/10.21437/Interspeech.2016-171.

Xu, L., Zhou, N., 2011. Tonal languages and cochlear implants. In: Zeng, F.G., Popper, A.N., Fay, R.R. (Eds.), Auditory Prostheses: New Horizons. Springer, New York, NY, pp. 341–364. https://doi.org/10.1007/978-1-4419-9434-9_14.

Xu, L., Xi, X., Patton, A., Wang, X., Qi, B., Johnson, L., 2020. A cross-language comparison of sentence recognition using American English and Mandarin Chinese HINT and AzBio sentences. Ear Hear., published online ahead of print. https://doi.org/10.1097/AUD.0000000000000938.

Yang, Z., Chen, J., Huang, Q., Wu, X., Wu, Y., Schneider, B.A., Li, L., 2007. The effect of voice cuing on releasing Chinese speech from informational masking. Speech

Commun. 49, 892–904. https://doi.org/10.1016/j.specom.2007.05.005.

Zekveld, A.A., Rudner, M., Kramer, S.E., Lyzenga, J., Rönnberg, J., 2014. Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. Front. Neurosci. 8, 88. https://doi.org/10.3389/fnins.2014.00088.

Zekveld, A.A., Koelewijn, T., Kramer, S.E., 2018. The pupil dilation response to auditory stimuli: current state of knowledge. Trends Hear 22. https://doi.org/10.1177/2331216518777174, 2331216518777174.

Zhang, J., Wang, X., Wang, N., Fu, X., Gan, T., Galvin, J.J., Fu, Q., 2020. Tonal language speakers are better able to segregate competing speech according to talker sex differences. J. Speech Lang. Hear. Res. 63 (8), 2801–2810. https://doi.org/10.1044/2020_JSLHR-19-00421.

Zurek, P.M., 1987. The precedence effect. In: Yost, W.A., Gourevitch, G. (Eds.), Directional Hearing. Springer-Verlag, New York.