Improving Speech Intelligibility Without Sacrificing Environmental Sound Recognition

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Eric M. Johnson, B.A., Au.D.

Graduate Program in Speech and Hearing Science

The Ohio State University

2022

Dissertation Committee

Eric W. Healy, Advisor

DeLiang Wang

Rachael Frush Holt

Abstract

The three manuscripts presented here examine concepts related to speech perception in noise and ways to overcome poor speech intelligibility without depriving listeners of environmental sound recognition.

Because of hearing-impaired (HI) listeners' auditory deficits, there is a substantial need for speech-enhancement (noise reduction) technology. Recent advancements in deep learning have resulted in algorithms that significantly improve the intelligibility of speech in noise, but in order to be suitable for real-world applications such as hearing aids and cochlear implants, these algorithms must be causal, talker independent, corpus independent, and noise independent. Manuscript 1 involves human-subjects testing of a novel, time-domain-based algorithm that fulfills these fundamental requirements. Algorithm processing resulted in significant intelligibility improvements for both HI and normal-hearing (NH) listener groups in each signal-to-noise ratio (SNR) and noise type tested.

In Manuscript 2, the range of speech-to-background ratios (SBRs) over which NH and HI listeners can accurately perform both speech and environmental recognition was determined. Separate groups of NH listeners were tested in conditions of selective and divided attention. A single group of HI listeners was tested in the divided attention experiment. Psychometric functions were generated for each listener group and task type.

It was found that both NH and HI listeners are capable of high speech intelligibility and high environmental sound recognition over a range of speech-to-background ratios. The range and location of optimal speech-to-background ratios differed across NH and HI listeners. The optimal speech-to-background ratio also depended on the type of environmental sound present.

Conventional deep-learning algorithms for speech enhancement target maximum intelligibly by removing as much noise as possible while maintaining the essential characteristics of the target speech signal. Manuscript 3 tests a new form of time-frequency masking that is designed to leave a small amount of background noise intact. The purpose of the unremoved background noise is to allow for environmental sound awareness while still providing significantly increased intelligibility. It was found that this type of processing resulted in significantly improved intelligibility and high environmental sound recognition performance for both types of listeners. It was also found that the same level of maximum attenuation provided the optimal balance of intelligibility and environmental sound recognition for both listener types.

Dedication

Dedicated to my wife, Kelly Ann, and our children.

Acknowledgements

Firstly, I would like to thank my advisor, Eric Healy, for providing me with exceptionally strong guidance and support these past five years. I owe my accomplishments to his expert mentorship and unwavering dedication to my success. I look forward to many more years of friendship and collaboration.

I would also like to extend a special thank you to my other dissertation committee members, DeLiang Wang and Rachael Frush Holt. I could not ask for more supportive encouragement. They were both integral to my growth during my doctoral career. I very much appreciate the mentorship and teaching expertise of Janet Weisenberger as well. I also treasure the wonderful friendships I have made with fellow students at Ohio State, including Victoria Sevich, Brittney Carter, Devan Lander, and Izabela Jamsek. Frederic Apoux and Jordan Vasko have also been excellent colleagues.

I gratefully acknowledge the encouragement and mentorship of Sarah Ferguson and Skyler Jennings at the University of Utah, who set me on the path to the Ph.D.

Finally, I want to thank my family, especially my wife, Kelly Ann. None of this would have been possible without her constant support. My love and endless gratitude to my

mother, Deirdra, my brother, Adam, my sister, Rachel, my father- and mother-in-law, David and Lana, and all of my brothers- and sisters-in-law. I also wish to extend a special thank you to three of my biggest fans, Fletcher, Gerrit, and Cordelia, who make every day better. And lastly, I want to acknowledge my late father, Kelly Lee Johnson, who showed me the importance of education and following my dreams.

Vita

June 24, 1988	Born
2013	B.A. Linguistics, Portuguese Brigham Young University Provo, UT
2013 – 2016	Graduate Teaching Assistant Department of Communication Sciences and Disorders University of Utah Salt Lake City, UT
2013 – 2016	Graduate Research Assistant Bruce L. Smith's Speech Science Laboratory Department of Communication Sciences and Disorders University of Utah Salt Lake City, UT
2016 – 2017	Audiology Extern VA Salt Lake City Health Care System Salt Lake City, UT
2017	Au.D. University of Utah Salt Lake City, UT
2018 – 2021	Graduate Teaching Associate Department of Speech and Hearing Science The Ohio State University Columbus, OH
2017 – 2022	Graduate Research Associate Eric W. Healy's Speech Psychoacoustics Laboratory Department of Speech and Hearing Science The Ohio State University Columbus, OH

Publications

Healy, E. W., Taherian, H., Johnson, E. M., & Wang D. L. (**2021**). "A causal and talker-independent speaker separation/dereverberation deep learning algorithm: Cost associated with conversion to real-time capable operation." J. Acoust. Soc. Am. **150**, 3976-3986.

Healy, E. W., Johnson, E. M., Delfarah, M., Krishnagiri, D. S., Sevich, V. A., Taherian, H., & Wang D. L. (**2021**). "Deep learning based speaker separation and dereverberation can generalize across different languages to improve intelligibility," J. Acoust. Soc. Am. **150**, 2526–2538.

Healy, E. W., Tan, K., Johnson, E. M., & Wang D. L. (**2021**). "An effectively causal deep learning algorithm to increase intelligibility in untrained noises for hearing-impaired listeners," J. Acoust. Soc. Am. **149**, 3943–3953.

Healy, E. W., Johnson, E. M., Delfarah, M., & Wang D. L. (**2020**). "A talker-independent deep learning algorithm to increase intelligibility for hearing-impaired listeners in reverberant competing talker conditions," J. Acoust. Soc. Am. **147**, 4106–4118.

Johnson, E. M., Morgan, S. D., & Ferguson, S. H. (**2020**). "Does time compression decrease intelligibility for female talkers more than for male talkers?" J. Speech, Lang. Hear. Res. **63**, 1083–1092.

Healy, E. W., Delfarah, M., Johnson, E. M., & Wang, D. L. (**2019**). "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," J. Acoust. Soc. Am. **145**, 1378–1388.

Smith, B. L., Johnson, E. M., & Hayes-Harb, R. (**2019**). "ESL learners' intra-speaker variability in producing American English tense and lax vowels," J. Sec. Lang. Pronun. **5**, 139–164.

Zhao, Y., Wang, D. L., Johnson, E. M., & Healy, E. W. (**2018**). "A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions," J. Acoust. Soc. Am. **144**, 1627–1637.

Hacking, J. F., Smith, B. L., & Johnson, E. M. (**2017**). "Utilizing electropalatography to train palatalized versus unpalatalized consonant productions by native speakers of American English learning Russian," J. Sec. Lang. Pronun. **3**, 9–33.

Johnson, E. M. (**2013**). "Spoken like a true poet: The recreation of speech in Manuel Bandeira's Libertinagem," La Marca Hispánica, **24**.

Fields of Study

Major Field: Speech and Hearing Science

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

An estimated 37.5 million Americans have hearing loss, and difficulty understanding speech in background noise is their most significant hearing handicap (Blackwell *et al.*, 2014). This is known as the "speech-in-noise" problem, and there is a substantial need to solve it. The primary treatment for hearing loss is hearing aids, and even though these devices can increase the audibility of low-level sounds, they fail to adequately improve speech understanding in noise. This is especially true when speech and noise are directionally co-located relative to the listener because directional microphone arrays can only provide benefit when there is spatial separation between sound sources. Restoring the ability to understand speech in noise, especially in environments where directional microphones are ineffective, will greatly enhance hearing-impaired (HI) listeners' quality of life by allowing them to better communicate in the many daily environments where background noise is present.

Fortunately, a groundbreaking technology called deep learning can effectively eliminate background noise from a sound mixture and substantially improve speech understanding for HI listeners. Wang and Wang (2013) introduced deep learning as a solution to the speech-in-noise problem, and Healy *et al*. (2013) provided the first demonstration of single-microphone noise reduction that can substantially improve intelligibility for HI listeners. For many listeners, this improvement was dramatic, and

several improved intelligibility from scores near zero to values above 70% correct. Deep-learning-based speech enhancement promises to revolutionize hearing aids and cochlear implants (CIs) by enabling them to overcome HI listeners' most significant auditory deficit.

Although groundbreaking at the time, the algorithm used in Healy *et al*., 2013, had three major limitations: (1) It was only trained to operate on a specific recording of a single talker; (2) it was designed to only work on two specific 10-s noise segments; and (3) it could not operate in real time. It is obviously impossible to train a deep neural network on every possible condition it may encounter during operation, making generalization to new conditions critical. Further, long processing delays are not tolerated by listeners (Stone and Moore, 1999; 2005; Goehring *et al*., 2018). Therefore, in order to have a direct translational impact, a speech-enhancement algorithm must generalize to talkers not used during training, including talkers from other speech corpora with different recording characteristics. This is known as talker/corpus independence. Second, it must also generalize to new noises. This is known as noise independence. Third, it must be real-time capable, only operating on current and past time frames. This is known as causality.

Manuscript 1 demonstrates the advances that have been made since the seminal study. It shows that a causal, talker-, corpus-, and noise-independent deep learning algorithm provides substantial intelligibility benefit to HI listeners. This algorithm used a time-domain-based scheme and complex representations to achieve high performance despite the great demands placed on it (Pandey and Wang, 2021). The performance of

this algorithm was also compared to that obtained in 2013. By using experimental conditions that were also used in the 2013 study, direct comparisons were made, and advancements were made clear.

However, deep-learning-based speech enhancement is so effective that it has had an unintended negative consequence – in some ways, we are victims of our own success. Unlike previous noise reduction approaches, that simply don't work very well and so substantial amounts of the background noise remains, deep learning can almost completely remove background sounds. So although it improves speech understanding dramatically, it also limits access to environmental sounds, which are an important aspect of the human auditory experience. Environmental sound awareness allows for greater personal independence and an improved sense of "connection" with one's surroundings (Harris *et al.*, 2017). The ability to perceive and respond to environmental sounds is also essential for personal safety and danger avoidance, and hearing loss is associated with an increased risk of injuries (Mick *et al.*, 2018). Many potential hazards are often forecast by environmental sounds, either naturally occurring or artificial. Examples of safety-relevant sounds include approaching vehicles, emergency sirens, vehicle horns, impact noises, and even the warning whistle of a lifeguard or crossing guard. The inability to detect and resulting failure to react to these environmental sounds could put HI individuals at risk for serious personal harm. Critically, therefore, noise reduction technology intended for hearing protheses should not unduly eliminate access to environmental sounds in the pursuit of maximizing intelligibility at all costs.

Manuscript 2 involves testing not only listeners' speech-in-noise recognition but also their "noise-in-speech" recognition. The purpose of this study was to determine the SNRs at which NH and HI listeners can recognize both speech *and* environmental sounds in conditions of selective attention and divided attention. Sentences were mixed with individual environmental sounds at several SNRs over a wide range, and speech intelligibility and environmental sound recognition (ESR) were measured. It was found that a range of SNRs exists over which both speech and background can be accurately received. The results of this study informed the design of Manuscript 3.

For many years, the primary goal of speech enhancement has been to maximize intelligibility in the presence of background noise by producing a signal that resembles clean speech as closely as possible. According to this philosophy, any background noise that remains after processing is a deviation from the desired outcome of perfectly uncorrupted speech. To my knowledge, Manuscript 3 represents the first study to test a time-frequency (T-F) mask that deliberately retains various levels of the background for the purpose of optimizing the trade-off between speech intelligibility and ESR. This modified version of the ideal ratio mask (IRM), known as the ideal compressed mask (ICM), increases intelligibility while preserving high levels of ESR. Like the standard IRM, the ICM attenuates T-F units on a sliding scale based on their relative levels of speech and noise, with noisier units being attenuated more and "cleaner" units, which are dominated by speech, being attenuated less. But whereas the IRM completely attenuates (i.e., zeroes out) any units with no speech energy, the ICM is limited in the amount of attenuation it can apply in order to limit the attenuation of the background. Hearing-

impaired and normal-hearing listeners were tested using ICMs with a range of different maximum-attenuation values, and the optimal value for jointly maximizing intelligibility and ESR was obtained. Even though this study did not involve mask estimation, previous work has shown that deep neural networks (DNNs) are capable of estimating the IRM with accuracy (Wang *et al*., 2014) sufficient to produce high speech intelligibility (e.g., Healy *et al.*, 2015, 2017, 2019; Chen *et al.*, 2016; Zhao *et al.*, 2018). Future studies will test the effect of the algorithm-estimated implementation of this training target based on the optimal value calculated in this study on speech recognition and ESR.

REFERENCES

Blackwell, D. L., Lucas, J. W., & Clarke, T. C. (**2014**). "Summary health statistics for U.S. adults: National Health Interview Survey," 2012. National Center for Health Statistics. Vital Health Stat, **10**. Retrieved from https://www.cdc.gov/nchs/data/series/sr_10/sr10_260.pdf

Chen, J., Wang, Y., Yoho, S. E., Wang, D. L., & Healy, E. W. (**2016**). "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," J. Acoust. Soc. Am. **139**, 2604-2612.

Goehring, T., Chapman, J. L., Bleeck, S., & Monaghan, J. J. (**2018**). "Tolerable delay for speech production and perception: Effects of hearing ability and experience with hearing aids," Int. J of Audio., **57**, 61-68.

Harris, M. S., Boyce, L., Pisoni, D. B., Shafiro, V., & Moberly, A. C. (**2017**). "The relationship between environmental sound awareness and speech recognition

skills in experienced cochlear implant users," Otology and Neurotology, **38**, e308–e314.

Healy, E. W., Delfarah, M., Johnson, E. M., & Wang, D. L. (**2019**). "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," J. Acoust. Soc. Am. **145**, 1378-1388.

Healy, E. W., Delfarah, M., Vasko, J. L., Carter, B. L., & Wang, D. L. (**2017**). "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," J. Acoust. Soc. Am. **141**, 4230-4239.

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., & Wang, D. L. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**, 1660-1669.

Healy, E. W., Yoho, S. E., Wang, Y., & Wang, D. L. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**, 3029-3038.

Mick, P., Foley, D., Lin, F., Pichora-Fuller, M. K. (**2018**). "Hearing difficulty is associated with injuries requiring medical care," Ear Hear. **39**, 631–644.

Pandey, A., & Wang, D. L. (**2021**). "Dense CNN with self-attention for time-domain speech enhancement," IEEE/ACM Trans. Audio. Speech Lang. Process. **29**, 1270-1279.

Stone, M. A., & Moore, B. C. (**1999**). "Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses," Ear Hear. **20**, 182-192.

Stone, M. A., & Moore, B. C. (**2005**). "Tolerable hearing-aid delays: IV. Effects on subjective disturbance during speech production by hearing-impaired subjects," Ear Hear. **26**, 225-235.

Wang, Y., & Wang, D. L. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio. Speech Lang. Process. **21**, 1381-1390.

Wang, Y., Narayanan, A., & Wang, D. L. (**2014**). "On training targets for supervised speech separation," IEEE/ACM Trans. Audio. Speech Lang. Process. **22**, 1849-1858.

Zhao, Y., Wang, D. L., Johnson, E. M., & Healy, E. W. (**2018**). "A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions," J. Acoust. Soc. Am. **144**, 1627-1637.

Chapter 2. Manuscript 1

Progress made in the efficacy and viability of deep learning based noise reduction

Eric W. Healy [a)], Eric M. Johnson

Department of Speech and Hearing Science, and Center for Cognitive and Brain Sciences

The Ohio State University, Columbus, OH 43210

Ashutosh Pandey, DeLiang Wang

Department of Computer Science and Engineering, and Center for Cognitive and Brain Sciences

The Ohio State University, Columbus, OH 43210

a) Electronic mail: healy.66@osu.edu

Running head: Implementing deep learning noise reduction

Abstract

Recent years have brought considerable advances to our ability to increase intelligibility through deep learning based noise reduction, especially for hearing impaired (HI) listeners. In the current study, steps taken toward creating an implementable algorithm are reviewed, and intelligibility improvements resulting from such an algorithm are assessed. Further, intelligibility benefits resulting from the current algorithm are compared to those resulting from the initial demonstration of deep learning based noise reduction for HI listeners (E.W. Healy *et al*., JASA, 134, 2013). The stimuli and procedures were essentially identical across studies. However, whereas the initial study involved highly matched training and test conditions, and non-causal operation, the current attentive recurrent network employed different talkers, speech corpora, and noise types for training versus test, and it was fully causal as required for real-time operation. Significant intelligibility benefit was observed in every condition, which averaged 51 percentage points across conditions for HI listeners. Further, benefit was comparable to that obtained in the initial demonstration, despite the considerable additional demands placed on the current algorithm. The retention of large benefit despite the systematic removal of various constraints as required for real-world operation reflects the substantial advances made to deep learning based noise reduction.

I. INTRODUCTION

Difficulty understanding speech in background noise remains the primary auditory complaint of hearing-impaired (HI) listeners. This problem persists despite the considerable technological advances made to hearing aids and other devices over several decades. A noise-reduction approach that has shown considerable promise involves deep learning. In this approach, a deep neural network (DNN) is trained to isolate target speech from various interferences including background noise, interfering speech, and/or room reverberation, allowing substantial increases in target-speech intelligibility. The current study was conducted to establish advances made toward implementing single-microphone deep learning noise reduction [1] into hearing devices to solve the speech-in-noise problem.

The implementation of deep learning based approaches into hearing technology can be divided into two broad considerations: efficacy and viability. The former refers to the ability of an algorithm to improve intelligibility for a wide variety of listeners (particularly HI), across a wide variety of acoustic environments. The latter involves the ability of an algorithm to operate in real time on an actual device. So whereas the former consideration involves the question, "does it work?," the latter involves the question, "can it work?"

Different approaches can be taken toward answering these questions and addressing the various challenges associated with each. The broad overall philosophy guiding our work in this area has been to focus first on efficacy. Accordingly, the early work involved algorithms focused more on large intelligibility improvements and less on

constraints associated with implementation. The initial work also involved limited sets of conditions, which were systematically expanded over various studies. By monitoring intelligibility benefit as the conditions were expanded and each step is taken toward viable implementation, performance "costs" associated with each step can be known and alternative approaches sought if that cost is high.

This approach contrasts with one adopted by others, which is essentially the opposite. In this opposite approach, work begins with a more real-world viable algorithm, and intelligibility improvements are sought. This approach results in smaller intelligibility improvements, but more problematically, this approach disguises the causes of the diminished performance.

In the sections below, an overview of progress made toward implementing deep learning based noise reduction is provided, followed by a description of a study that establishes intelligibility performance once all these advances are incorporated.

A. Efficacy considerations

The question, "does it work?" necessarily means, "does the algorithm increase target-speech intelligibility across a large variety of environmental conditions for a wide variety of HI listeners?" The key to efficacy is the ability of an algorithm to generalize to conditions not encountered during network training – to tolerate a training-test mismatch. This is critical because it is obviously impossible to train a network on all conditions that will be encountered by a listener in the real world.

In accord with our overall philosophy, the earliest work in the area employed highly matched training and test conditions, and large intelligibility increases were sought for HI listeners (Healy *et al*., 2013; Wang and Wang, 2013). This goal was met, with HI listeners displaying intelligibility increases averaging over 50 percentage points in the lower SNR conditions. Favorably, everyday sentences were employed, both stationary and nonstationary background noises were used, and various SNRs were tested. Obviously, different sentences were used for training versus test. But notably, the same talker and speech recordings were used for training and test (the network was talker dependent), and the same 10-s segment of nose was used for training and test (the network was noise dependent). Additionally, the network was trained on speech mixed with noise at the same SNR as that used for the test mixtures.

The path that followed this initial study involved systematically increasing the training-test mismatch – increasing the requirement that the algorithm generalize – while maintaining large intelligibility increases for HI listeners. The first step taken involved training using one segment of noise and testing using a different segment of the same noise – untrained noise segments (Healy *et al*., 2015). In order to accomplish this generalization, the training noise duration was expanded considerably and a perturbation technique (Chen *et al*., 2015) was used to expand the training set. The noise type was also expanded to include a real environmental recording containing multiple sound sources (cafeteria noise, which contains speech babble, impact sounds from dishes, etc.). Despite the increased challenge posed by this generalization and more complex noise, HI

listener intelligibility increases were comparable to those observed in the initial study in the comparable noise type/SNR.

The next step involved generalization to entirely different noises (Chen *et al*., 2016). This was accomplished by training on 10,000 different noises, in order to teach the network what "noise" was in a general sense, then testing on two completely different noise recordings not in the training set. This large-scale multi-condition training resulted in 640,000 training mixtures, each containing one of 560 training sentences mixed with a random segment of a random training noise. There was also a mismatch between training and test SNR for the HI listeners. The HI intelligibility benefits resulting from algorithm processing in Chen *et al*. (2016) were reduced somewhat relative to the more matched training-test conditions of Healy *et al*. (2015), but they were again observed despite the considerable generalization challenge.

Because efficacy also requires effective operation across a large variety of environmental conditions, additional corruptions were also examined. Individuals with hearing impairment have particular difficulty understanding speech in the presence of an interfering talker. In Healy *et al*. (2017), these conditions were examined because they represent a considerably different problem for human and machine listeners. A DNN was trained to separate a target talker from an interfering talker, and the target sentence was passed along to the listener (speaker separation). Intelligibility benefit for HI listeners was found to be considerable, averaging 50 percentage points across conditions.

Another corruption common to everyday acoustic environments involves room reverberation. Room reverberation is often highly detrimental to speech perception,

especially for HI listeners, and especially when combined with background noise or interfering speech. However, reverberation and additive noise corrupt the acoustic speech signal in very different ways, and so the presence of the concurrent corruptions represents a substantial challenge for an algorithm designed to increase intelligibility.

Zhao *et al*. (2018) examined the ability of a deep learning algorithm to improve intelligibility for HI listeners in the presence of background noise and concurrent room reverberation. Target sentences were subjected to room reverberation with a T_{60} value of 0.6 s and mixed with one of two noise types at various target-to-interferer ratios (TIRs). By targeting the noise-free reverberation-free speech, the network performed simultaneous noise reduction and de-reverberation. Significant intelligibility benefit was observed for HI listeners in all conditions.

Healy *et al*. (2019) examined interfering speech and concurrent room reverberation and addressed whether greater intelligibility benefit was obtained for HI listeners when the network was trained to remove interfering speech and reverberation or only remove interfering speech and allow reverberation to remain. Whereas no difference in benefit was observed across these conditions for normal-hearing (NH) listeners, greater intelligibility benefit was observed for the HI listeners (56 percentage points in the lowest TIR condition) when simultaneous noise reduction and de-reverberation was performed.

Another aspect of generalization that is critical for real-world efficacy involves talker independence – that an algorithm be able to increase intelligibility of any voice the listener encounters. [2] This generalization was introduced to conditions involving

interfering speech and concurrent room reverberation, and intelligibility benefit was assessed for HI listeners by Healy *et al*. (2020). Network training was conducted using 101 different talkers and testing was performed using a pair of talkers from a different speech corpus. The algorithm operated in the complex domain (Williamson *et al*., 2016), in which both the amplitude and phase of the target speech was estimated from the speech-plus-interference signal, rather than estimating amplitude alone, which is typically done. Considerable benefit was observed for HI listeners in all conditions.

The use of highly similar conditions across Healy *et al*. (2019) and Healy *et al*. (2020) allows direct comparison. Both studies employed reverberant single-talker interference, the same speech materials, and a condition involving TIR = 0 dB, T_{60} = 0.6 s. The use of a different algorithm and complex representations in the latter study allowed greater benefit to be observed (73 percentage points) compared to the corresponding condition of the former study (56 percentage points), despite the additional challenge associated with talker independence in the latter study.

An additional generalization involves the use of different speech corpora for network training versus testing, particularly when these corpora are recorded using different equipment in different environments. Whereas human listeners adapt quickly and easily to these differences, machines can be more sensitive to the constant transfer-function characteristics associated with a fixed recording environment. In practice, the use of a large multi-talker corpus for training and a more standardized corpus for testing produces not only talker independence but also corpus/recording channel independence (e.g., Healy *et al*., 2020). Pandey and Wang (2020a) attributed the challenge of cross-

corpus generalization largely to differences in recording channels and proposed techniques to address cross-corpus generalization.

Most recently, the concept of generalization to conditions not employed during network training was pushed to perhaps a limit (Healy *et al*., 2021a). A DNN was used to isolate a target talker from an interfering talker and simultaneously remove large amounts of room reverberation. Training was performed using English speech materials, but testing was performed using Mandarin speech materials. These two languages were selected based on their high prevalences of speakers and their lack of known common ancestry and correspondingly large linguistic differences. Additional generalizations included speech corpus/recording channel, target-to-interferer energy ratios, reverberation room impulse responses, and test talkers.

Despite that only normal-hearing (NH) listeners were tested, large intelligibility increases were observed, which averaged 44 percentage points across conditions. Further, the benefit observed in cross-language conditions were comparable to those observed in within-language conditions, suggesting that vast generalizations are possible and that network performance was not hindered by the challenge associated with generalizing to an entirely different language. This work further suggests that the learning that takes place by the DNN transcends language and is instead perhaps more centered on aspects of the particular speech sounds that humans can produce.

A. Viability considerations

The question, “can it work?” necessarily means, “can the trained network be implemented into mobile technology, operate in real-time, and retain the ability to produce large intelligibility benefits?” There is only one fundamental requirement for viability – causality – that a network operate on only past and present time frames and not delay output in order to take advantage of future time frames.

In accord with our broad philosophy involving a focus on network performance (the ability to provide large intelligibility benefit in a variety of scenarios), the papers described thus far involved networks that employed large analysis windows including both past and future time frames. The advantage to this approach is obvious, as changes in spectro-temporal speech energy can be predicted by both past and future events, which result from gradual changes in speech-articulator position. But the need to eliminate the use of future time frames is also obvious, as this introduces a time delay to the processed (noise-reduced) signal.

The causality requirement was met by Healy *et al*. (2021b) through the concept of “effectively causal.” Future time frames that produce delays below the human detection or disruption threshold may be used without hinderance to the listener but with potential benefit to the deep learning algorithm. A network that operated in the complex domain was used to remove complex noises from speech. Significant intelligibility increases were observed for HI listeners in all of the effectively causal conditions.

The “cost” associated with transforming a DNN to causal operation was examined by Healy *et al*. (2021c). The network employed by Healy *et al*. (2020) was modified to be fully causal and the isolation of a target talker from complex interference involving

competing speech and concurrent room reverberation was assessed. Intelligibility benefit for HI listeners remained high (averaging 47 percentage points across conditions), despite the removal of future time frames. A decrement in intelligibility benefit resulting from conversion to causal operation was present in most but not all conditions, and these decrements were statistically significant in half of the conditions tested. It was concluded that a cost associated with causal processing was present in most conditions, but may be considered modest relative to the overall level of benefit.

The other main requirement for viability involves the size/efficiency of the network and its ability to operate on small mobile platforms. Unlike causality, this aspect is not fundamental because it is relative to the processing capability of the platform, which is increasing dramatically each year. But it is clear that smaller/more efficient networks are more readily implementable. Those who have adopted the opposite approach described above have shown that even small networks having reduced computational complexity can increase intelligibility. Monaghan *et al*. (2017) demonstrated significant intelligibility benefits for HI listeners using small and causal noise-reduction neural networks using a matched training-test speech corpus and untrained segments of noises. Keshavarzi *et al*. (2019) showed that HI listeners displayed a slight subjective intelligibility preference for speech extracted from multi-talker babble using a small, causal, and talker-independent neural network. Goehring *et al*. (2017) and Goehring *et al*. (2019) employed small, causal, talker-independent neural networks and observed improved intelligibility in noise for cochlear-implant (CI) users in some noise types but not others.

Real-time deep learning based noise reduction has recently been implemented into commercial products, which demonstrates the real-world feasibility of such applications. For example, the software plugin Krisp (https://krisp.ai/) uses artificial neural networks to remove noise in voice and videoconferencing applications such as Zoom. However, its capacity to increase intelligibility, especially for HI listeners, has yet to be determined to our knowledge. More relevant for listeners with hearing loss, a hearing aid containing on-board deep learning based noise reduction was made available in 2021 – the Oticon More™ line. Using speech in diffuse noise, intelligibility increases of approximately 5 percentage points and 1.2 to 1.5 dB speech reception threshold were observed relative to the prior generation Oticon hearing aid (Santurette *et al*. 2020).

In the current study, HI and NH intelligibility benefit resulting from a current deep learning noise-reduction algorithm was assessed. These results were then compared to that associated with the initial study (Healy *et al*., 2013). To allow direct comparison, the speech recordings, noise types, some SNRs, test procedures, and listener populations were the same across studies. However, whereas the initial study involved highly similar training and test conditions (required little generalization), the network employed currently was tasked with extensive generalization. The initial study employed the same talker and sentence corpus for training and test (it was talker and corpus dependent) and the same noises for training and test (it was noise dependent). In contrast, the current network was trained using over 2,000 different talkers and tested using a talker not included in this set and from an entirely different corpus (talker and corpus independent). Further, it was trained using 10,000 noises and tested on entirely different noises (noise

independent). Finally, whereas the initial algorithm employed an analysis time window involving both past and future time frames (non-causal), the current network was fully causal as required for real-time operation. The overall goal of the current study was to establish intelligibility benefit lost/gained across the span of steps taken toward the creation of an implementable deep learning based noise-reduction algorithm for mobile hearing technology.

II. METHOD

A. Subjects

Two groups of listeners participated. The HI group consisted of 12 listeners, representing typical hearing aid users with bilateral sensorineural hearing loss. All were binaural hearing aid users, and seven were female. These HI listeners were recruited from The Ohio State University Speech-Language-Hearing Clinic and ranged in age from 20 to 85 years (mean = 57). Pure-tone audiometry (ANSI, 2004, 2010) was used to verify hearing losses on day of test. In accord with our desire to recruit representative audiology patients, the degree of hearing loss varied across frequencies and listeners, who generally had sloping hearing losses that ranged in degree from mild to profound. Pure-tone average audiometric thresholds (PTAs; means across thresholds at 500, 1000, and 2000 Hz across ears) ranged from 27 to 76 dB hearing level (HL) with a mean of 56 dB HL. Three of the listeners had audiometric thresholds within normal limits (20 dB HL or lower) for at least one frequency in at least one ear (see dashed horizontal line in Fig. 1), but otherwise all thresholds were elevated in both ears for all audiometric frequencies.

Figure 2.1 displays audiograms for each of these listeners, who are numbered from HI1 to HI12 in order of ascending PTA.

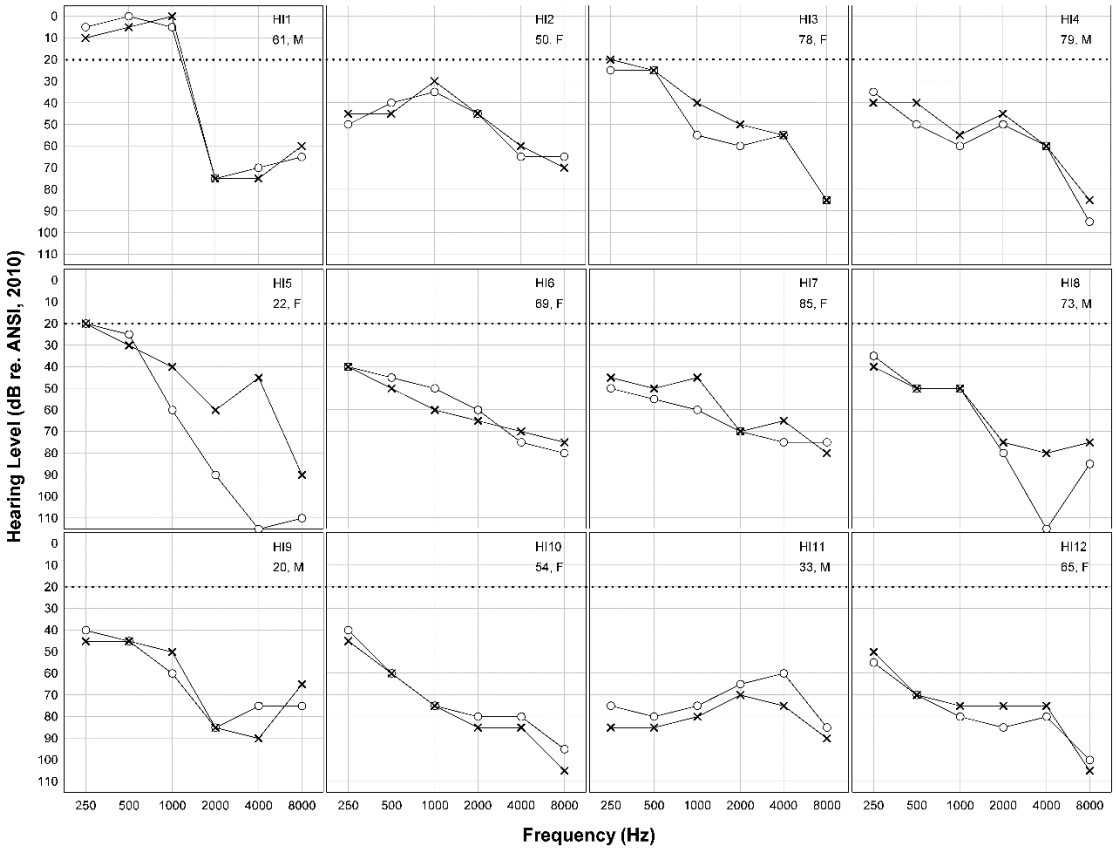

Figure 2.1. Pure-tone air-conduction audiograms for the HI listeners. Circles represent right ears and ×'s represent left ears. The horizontal dotted line in each panel represents the NH limit of 20 dB HL. Listener ages and sexes are also listed.

The NH group consisted of 12 listeners (all female) with pure-tone audiometric thresholds of 20 dB HL or lower at octave frequencies from 250 to 8000 Hz on day of test (ANSI, 2004; 2010). The exception was NH10, whose threshold at 8,000 Hz in the right ear was 25 dB HL. Recruited from undergraduate courses at The Ohio State University, they ranged in age from 19 to 26 years (mean = 21) and represented young

listeners with "ideal" hearing abilities. All participants (HI and NH) were native speakers of American English having no previous exposure to the test sentences used in the current study, and all received either extra course credit or a monetary incentive for participating.

B. Stimuli

Target sentences were drawn from the standard recordings of the Hearing in Noise Test (HINT; Nilsson *et al*., 1994) and were all produced by the same male talker in general American English. Speech-shaped noise (SSN) and multi-talker babble were used as background noises. The babble was a standard recording from auditec (http://www.auditec.com). HI listeners were tested at -2 and -5 dB signal-to-noise ratio (SNR) in SSN and at 0 and -2 dB SNR in babble. The NH listeners were tested at -2 and -5 dB SNR in both SSN and babble. Stimuli were presented as unprocessed noisy sentences and as algorithm-processed (enhanced) versions of noisy sentences.

The algorithm was trained using speech materials from the Librispeech corpus (Panayotov *et al*., 1992). Librispeech is a corpus of approximately 1,000 hours of speech from more than 2,000 speakers. It is primarily used for research on large-vocabulary continuous speech recognition systems. The data in Librispeech are derived from the LibriVox project (Kearns, 2014), which contains audiobook recordings created using volunteers from across the globe. Pandey and Wang (2020a, 2020b) found Librispeech to be a highly effective corpus for training corpus-independent speech enhancement algorithms. Since Librispeech is recorded by thousands of volunteers in diverse environments, recording conditions vary considerably within the dataset, which is a key

to avoiding overfitting on corpus specific characteristics, such as recording microphones and room acoustics.

The training noises consisted of 10,000 nonspeech sounds from a sound-effect library (Richmond Hill, ON, Canada[3]). Pairs of clean and noisy speech were created during training by randomly selecting an utterance, a noise segment, and an SNR from {-5, -4, -3, -2, -1, 0} dB. A set of 150 validation mixtures was generated using utterances from 6 speakers in the Wall Street Journal Corpus (WSJ0; Paul and Baker, 1992) and a factory noise from the NOISEX dataset (Varga and Steeneken, 1993).

C. Algorithm description

Given a speech signal $\boldsymbol{s}$ and a noise signal $\boldsymbol{n}$ with N samples, a noisy speech signal $\boldsymbol{y}$ is modeled as

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n} \quad (1)$$

, where $\{y, s, n\} \in \mathbb{R}^{1\times N}$. A speech enhancement algorithm is concerned with improving the intelligibility and quality of noisy speech $\boldsymbol{y}$ by obtaining a good estimate of $\boldsymbol{s}$ (i.e., $\hat{\boldsymbol{s}}$) from $\boldsymbol{y}$. The current model was a time-domain algorithm. For such an algorithm, the input feature is the time-domain signal $\boldsymbol{y}$ instead of a time-frequency representation, such as STFT, and the estimated output is the time-domain signal $\hat{\boldsymbol{s}}$. This approach alleviates the need to perform fast Fourier transformations then the inverse transforms to return to the time domain.

In the current algorithm, an attentive recurrent network (ARN) was employed for mapping a noisy waveform to an enhanced waveform (Pandey and Wang, 2022). The

ARN accepts speech-plus-noise input and so is a monaural (single-microphone) system. The algorithm is made causal by restricting the use of time-frame information to the current and past frames. A block diagram of the ARN for time-domain speech enhancement is shown in Fig. 2.2.

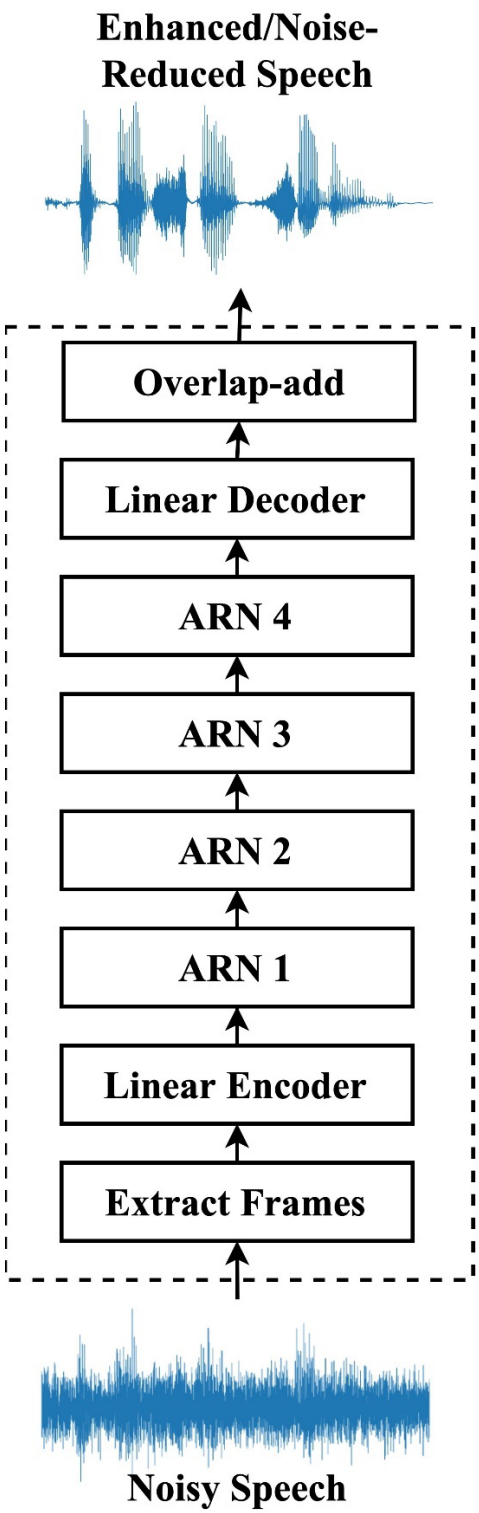

Figure 2.2. The employed ARN for time-domain speech enhancement.

The current algorithm represents a major advance relative to the initial study on intelligibility improvements of noisy sentences for HI listeners (Healy *et al*., 2013). First, and as mentioned in Sec. I, the model of Healy *et al*. (2013) was speaker and noise dependent, in which the training utterances were a subset of the test HINT sentences, and training and test noises were selected from the same 10-s segment of noise. In contrast, the current model is corpus, speaker and noise independent. The Librispeech training utterances are very different from the HINT test sentences and the 10,000 noises used for training also differ from those used for testing. It is noted that developing a corpus-independent DNN in low SNR conditions has been found to be particularly challenging due to the recently revealed cross-corpus generalization issue in DNNs (Pandey and Wang, 2020a).

Second, the current algorithm represents a large improvement in terms of speech enhancement problem formulation. In the initial study, speech enhancement was formulated as magnitude-only enhancement and trained subband DNNs were used to estimate the ideal binary mask (IBM). The goal of the algorithm was to obtain an accurate estimate of the IBM and not the clean speech. This formulation requires the use of noisy phase for reconstruction but also leads to limited magnitude enhancement. The current study formulates speech enhancement in the time domain, where the goal is to directly estimate the enhanced waveform from the noisy waveform, and as a result, the magnitude and the phase are jointly enhanced. In the case of ideal estimation, a time-

domain algorithm will output the clean speech, whereas the IBM will output intelligible speech but with variable quality. Moreover, a time-domain algorithm does not require feature extraction at the input or waveform reconstruction at the output. Healy *et al*. (2013) used a set of complementary features (from Wang *et al*., 2013) at the input and performed waveform reconstruction at the output using gammatone filterbanks.

Finally, the current algorithm represents an advance in terms of model architecture. It employs modern DNN building blocks suitable for efficient sequential processing with contextual information, such as long short-term memory (LSTM; Hochreiter and Schmidhuber, 1997) and self-attention (Vaswani *et al*., 2017). In contrast, the initial study employed multilayer perceptrons with pretraining based on restricted Boltzmann machines (see Wang and Wang, 2013).

In the ARN currently employed for speech enhancement, a noisy input $\boldsymbol{y}$ is first segmented into overlapping frames using a frame size of L samples and a frame shift of H samples to get $\boldsymbol{Y} \in \mathbb{R}^{T \times L}$, where T is the number of frames. Next, all the frames are projected to a higher dimension of size D using a linear encoder. The output from the encoder is processed using a stack of four ARNs. The output from the final ARN is projected back to the original frame size of L using a linear decoder at the output. Finally, overlap-and-add is applied to the sequence of enhanced frames to obtain the enhanced waveform.

The building blocks of the ARN are shown in Fig. 2.3. The ARN is composed of a recurrent neural network (RNN) block, an attention block, and a feedforward block. The input to the RNN block is first normalized using a layer normalization (Ba *et al*.,

2016) and then processed using an LSTM RNN. Layer normalization is used for faster training convergence and improved generalization. LSTM is used to model the temporal dependency between the sequence of frames in a causal fashion. The input and the output of the RNN block are of size $T \times D$.

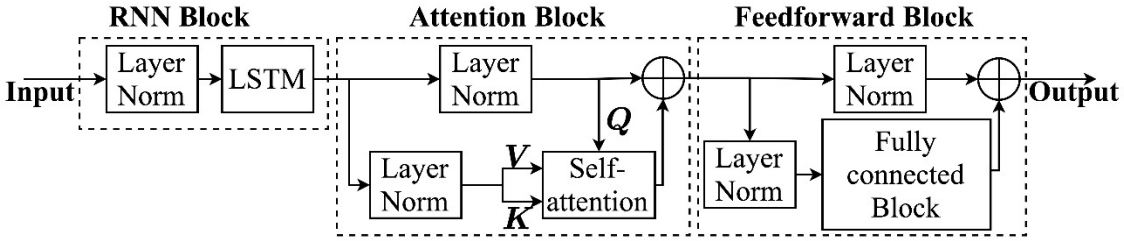

Figure 2.3. The building blocks of the ARN. It comprises an RNN block, an attention block, and a feedforward block. Layer Norm denotes layer normalization and $\oplus$ is the elementwise addition operator.

The RNN block is followed by the attention block. The input to the attention block is normalized using two separate layer normalizations having different scale and bias parameters. The first output is used as query ($\boldsymbol{Q}$), and the second output is used as key ($\boldsymbol{K}$) and value ($\boldsymbol{V}$) for a following self-attention mechanism, where $\{\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\} \in \mathbb{R}^{T \times D}$.

A schematic of the self-attention mechanism is shown in Fig. 2.4. It involves three trainable vectors $\{\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}\} \in \mathbb{R}^{1 \times D}$. The rows in $\boldsymbol{Q}, \boldsymbol{K}$, and $\boldsymbol{V}$ are refined using the following gating mechanism.

$$\boldsymbol{K}' = \boldsymbol{K} \odot \sigma(\boldsymbol{k}) \tag{2}$$

$$\boldsymbol{Q}' = \mathrm{Lin}(\boldsymbol{Q}) \odot \sigma(\boldsymbol{q}) \tag{3}$$

$$\boldsymbol{V}' = \boldsymbol{V} \odot \left[\sigma(\mathrm{Lin}(\boldsymbol{v})) \odot \mathrm{Tanh}(\mathrm{Lin}(\boldsymbol{v}))\right] \tag{4}$$

, where σ is the sigmoidal nonlinearity, Lin is a linear layer and $\odot$ is elementwise multiplication. Before the elementwise multiplication, vectors $\boldsymbol{q}$, $\boldsymbol{k}$, and $\boldsymbol{v}$ are broadcasted to match with the shape of matrices $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ respectively. Given that $\boldsymbol{v}$ is a fixed vector, $\sigma(\mathrm{Lin}(\boldsymbol{v})) \circledcirc \mathrm{Tanh}(\mathrm{Lin}(\boldsymbol{v}))$ is also a fixed vector. This operation is only required at training time for optimization of $\boldsymbol{v}$ (Merity, 2019). A precomputed constant vector from the best model after training is used during evaluation.

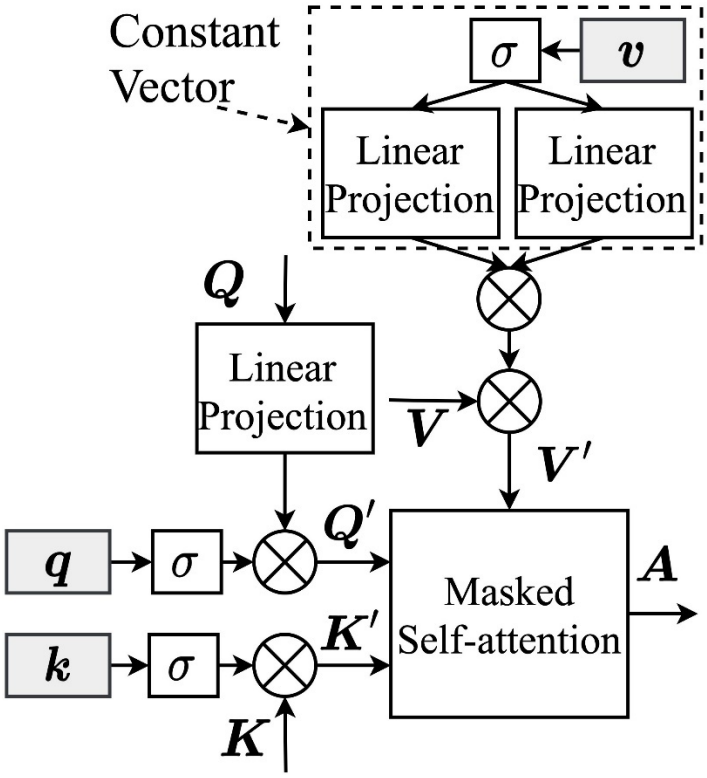

Figure 2.4. The self-attention mechanism inside the attention block. The inputs to self-attention are Q,K and V, and the final output is A. Vectors q,k, v and parameters of linear projections are trainable.

The final output of the attention block, $\boldsymbol{A} \in \mathbb{R}^{T \times D}$, is computed using the following set of equations.

$$W = \frac{\boldsymbol{Q}'\boldsymbol{K}'^{\mathbb{T}}}{\sqrt{D}} \tag{5}$$

$$\boldsymbol{W}' = \text{Mask}(\boldsymbol{W}) \tag{6}$$

$$W'(i,j) = \begin{cases} W(i,j), & \text{if } i \leq j \\ -\infty, & \text{otherwise} \end{cases} \tag{7}$$

$$\boldsymbol{P} = \text{Softmax}(\boldsymbol{W}') \tag{8}$$

$$\text{Softmax}(W)(i,j) = \frac{e^{W(i,j)}}{\sum_{j=1}^{T} e^{W(i,j)}} \tag{9}$$

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{V}' \tag{10}$$

, where $\mathbb{T}$ is the transpose operator.

First, correlation scores between pairs of rows in $\boldsymbol{Q}'$and $\boldsymbol{K}'$, $\{\boldsymbol{Q}'_{\boldsymbol{i}}, \boldsymbol{K}'_{\boldsymbol{j}}\}$, where $i, j \in \{1, \cdots T\}$, are computed using Eq. (5). Next, correlation scores of future frames are masked or ignored by using a mask operator defined in Eq. (7), which sets the correlation scores of future frames to $-\infty$. Then, the softmax operator in Eq. (9) is applied to convert correlation scores to probability values $\boldsymbol{P} \in \mathbb{R}^{T \times D}$. The softmax operator uses exponential followed by summation in the denominator. The exponential converts a $-\infty$ (from mask) to 0, and hence, the contribution of future frames in the total sum of denominator becomes zero, which makes the algorithm fully causal. Finally, the attention output, $\boldsymbol{A} \in \mathbb{R}^{T \times D}$, is computed using Eq. (10). The final output from the

attention block is obtained by adding $\boldsymbol{A}$ to $\boldsymbol{Q}$, which provides a residual connection for improved gradient flow during training (He *et al*., 2016).

The output from the attention block is processed using a feedforward block. The feedforward block provides additional representation power to the preceding attention block (Vashwani *et al*., 2017). The input to the feedforward block is normalized using two separate layer normalizations. The first normalized input is processed using a fully connected block shown in Fig. 2.5. In the fully connected block, a linear layer is first used to project its input of size D to a higher dimension of size $4D$, which is followed by Gaussian error linear unit (GELU) nonlinearity (Hendrycks and Gympel, 2016) and dropout. The output of size $4D$ is then collapsed to size D by splitting it into 4 different vectors and adding them together. Finally, the collapsed output is added to the second normalized input to get the final output of the feedforward block.

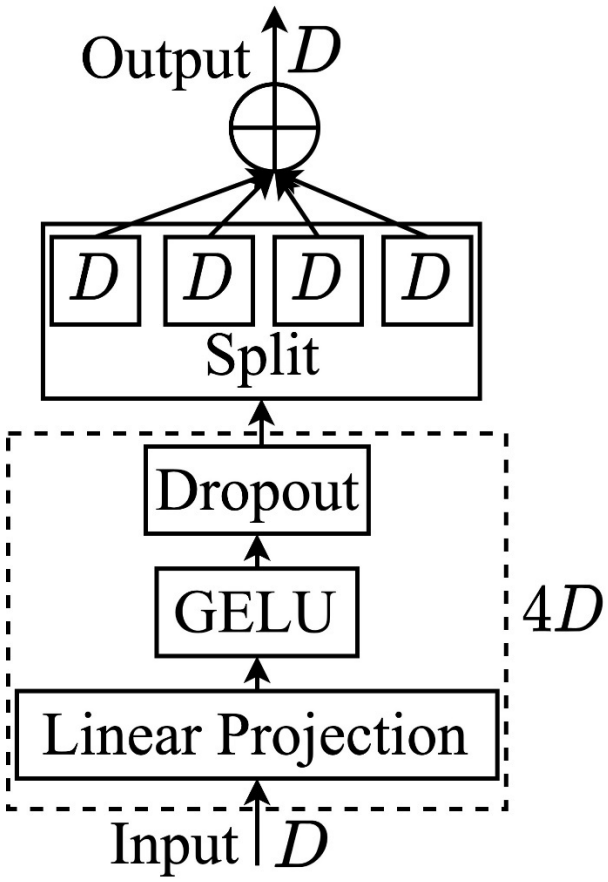

Figure 2.5. The fully connected block inside the feedforward ARN block.

All stimuli were resampled to 16 kHz for processing. Prior to mixing, the target (clean) speech was scaled to achieve the desired SNR. Next, the input to the network was RMS normalized. A frame size of 20 ms (L = 320) and a frame shift of 2 ms (H = 32) was used. The use of a smaller frame shift was inspired by earlier studies on corpus-independent speech enhancement (Pandey and Wang, 2020a; Pandey and Wang, 2020b), where a smaller frame shift led to improved generalization on untrained corpora. The RNN block used LSTM with hidden size of 1024. The parameter D was set to 1024. A dropout of 5% was used in the fully connected block.

The ARN was trained for 100 epochs with a batch size of 32 utterances. The pairs of clean and noisy utterances were dynamically generated during training by adding random segments of speech to random segments of noise. All the utterances within a batch were either truncated or padded with zeros to have length of 4 seconds.

The Adam optimizer was used for training (Kingma and Ba, 2014). The learning rate was set to 0.0002 for the first 33 epochs after which it is exponentially decayed every epoch using a scale that resulted in a learning rate of 0.00002 in the final epoch. All the models were developed in PyTorch. Mixed precision training was used to increase efficiency (Micikevicius *et al*., 2017). The ARN was trained using two Nvidia V100 32GB GPUs, with each batch distributed over two GPUs using PyTorch's DataParallel module.

Figure 2.6 displays waveforms and spectrograms for clean, noisy, and algorithm-processed versions of a HINT sentence used in the present study.

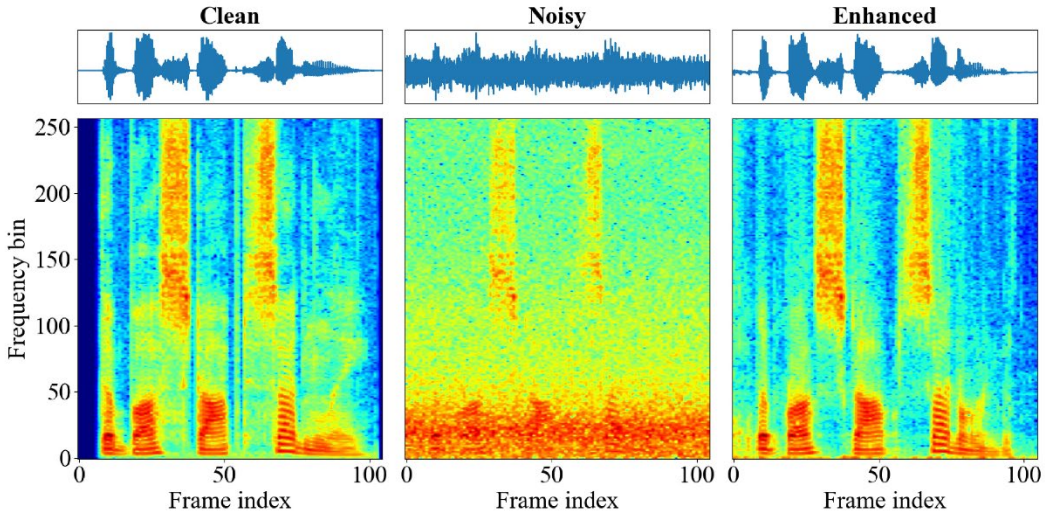

Figure 2.6. Enhancement of a HINT utterance corrupted by speech-shaped noise at -5 dB SNR using the employed ARN.

D. Procedure

There were eight conditions for each listener (two processing conditions × two noise types × two SNRs). Each listener heard 160 sentences, blocked by condition, with 20 sentences in each block. Within each combination of noise type and SNR, unprocessed and processed conditions were presented juxtaposed. The presentation order of the four noise type-SNR combinations was randomized for each listener, as was the order of the two processing conditions within each noise type-SNR block. The sentence materials were presented to each listener in a fixed order to ensure a random correspondence between sentences and conditions. No sentence was used more than once for any listener.

The stimuli were played back from a Windows PC using an RME Fireface UCX digital-to-analog converter (Haimhausen, Germany), through a Mackie 1202-VLZ mixer (Woodinville, WA), and presented diotically using Sennheiser HD 280 Pro headphones (Wedemark, Germany). The overall RMS level of each stimulus was set to 65 dBA in each ear using a sound-level meter and flat-plate coupler (Larson Davis models 824 and AEC 101, Depew, NY). For the HI listeners, additional frequency-specific gains were applied to compensate for the hearing loss of each individual listener using the NAL-RP hearing-aid fitting formula (Byrne *et al*., 1990). This formula does not prescribe gains at 125 or 8000 Hz, and so the gains applied to 250 and 6000 Hz (respectively) were also applied to these two most extreme standard audiometric frequencies. These gains were implemented using a RANE DEQ 60 L digital equalizer (Mukilteo, WA), as described in Healy *et al*. (2015). Accordingly, these listeners were tested without their hearing aids.

Twenty-five practice stimuli were presented to each listener prior to formal testing, consisting of five stimuli in each of the following five conditions: (1) sentences in quiet, (2) processed sentences in babble at the higher of the two SNRs for each listener group, (3) processed sentences in SSN at the lower SNR for each listener group, (4) unprocessed sentences in babble at the higher SNR, and (5) unprocessed sentences in SSN at the lower SNR. During this familiarization, listeners were instructed to repeat back each sentence as best they could, and guess if unsure of the content of the sentence. Hearing-impaired listeners were also asked to report about the loudness of the signals. All but two of the HI listeners reported that the stimuli sounded audible and comfortable. HI10 reported that they sounded comfortable after reducing the presentation level by 5 dB. HI12 reported that the stimuli sounded comfortable and audible after a 10-dB reduction in presentation level. The final presentation level for the HI listeners ranged from 81.0 to 98.6 dBA (mean = 91.2 dBA).

Listeners then heard the 160 test stimuli while seated in a double-walled sound booth. They were again instructed to repeat back each sentence to their best ability, guessing if unsure. The experimenter controlled the presentation of each stimulus and scored words correctly reported. For a word to be scored as correct, it had to be repeated exactly apart from verb tense (is/was, are/were, and has/had) and article (a/the) variations. Each of the 20 target sentences presented in each condition each contained between three and seven words, for a total of 103 to 110 words in each condition. Sentence recognition was expressed as the percentage of words correctly reported, and

these percent-correct scores were transformed to rationalized arcsine units (RAUs; Studebaker, 1985) prior to statistical analysis.

III. RESULTS AND DISCUSSION

A. HI listeners

Figures 2.7 and 2.8 display intelligibility scores for each individual HI listener in each condition. Results for the SSN conditions are displayed in Fig. 2.7, and those for the babble conditions are displayed in Fig. 2.8. Each panel corresponds to a different SNR, which is indicated. The black and shaded/hatched columns represent scores before and after algorithm processing, respectively. The absence of a black column for HI12 in SSN at -5 dB SNR reflects that she was unable to correctly report any words in that unprocessed condition. The algorithm benefit for each listener in each condition corresponds to the difference in height between a shaded/hatched column and the black column immediately to the left of it.

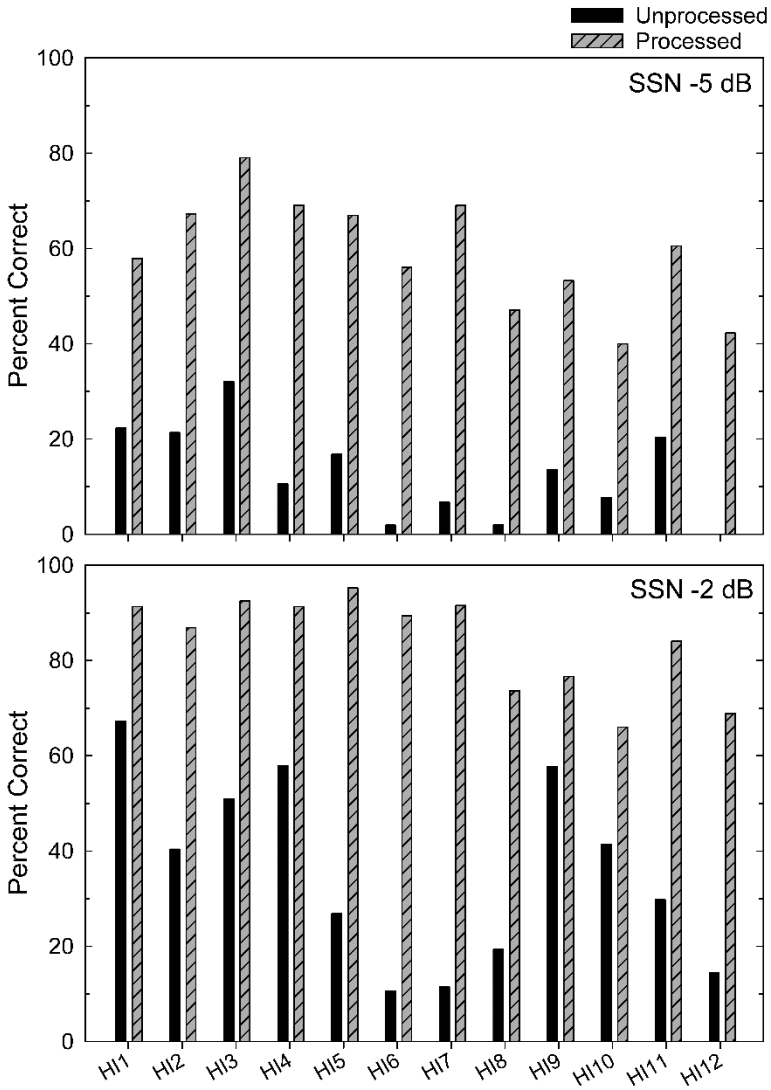

Figure 2.7. Sentence intelligibility scores for individual HI listeners. Black columns represent unprocessed speech-in-noise conditions and shaded/hatched columns represent these same conditions following algorithm processing. Algorithm benefit is then represented as the difference between a shaded/hatched column and the solid column directly to its left. The background noise type and SNR are indicated.

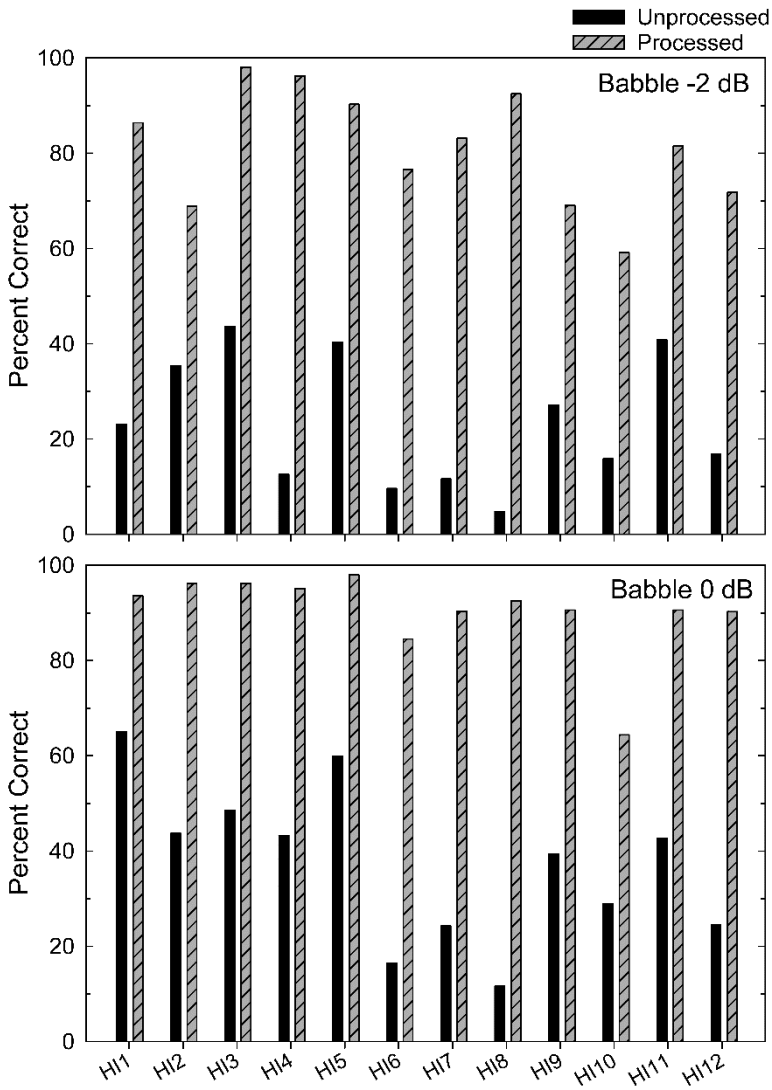

Figure 2.8. As Fig. 2.7, but for individual HI listeners and speech in multitalker babble.

Apparent from Fig. 2.7 is that the algorithm benefitted all HI listeners at both SNRs in SSN. At least half of the HI listeners received benefit exceeding 45 and 50 percentage points for the SNRs of -5 and -2 dB, respectively. Algorithm benefit in SSN exceeded 30 percentage points in 21 out of the 24 cases (12 HI listeners × 2 SNRs). Apparent from Fig. 2.8 is that all HI listeners also received benefit at both SNRs in

babble, with at least half of them receiving benefit exceeding 50 percentage points at both SNRs. The benefit in babble exceeded 30 percentage points in 23 of the 24 cases.

Planned comparisons consisting of four two-tailed paired *t*-tests on RAUs revealed significant algorithm benefit for HI listeners at each of the SNRs tested in both noise types [each $t(11) \geq 7.9$, each two-tailed p value < 0.0001]. These significant results survive Bonferroni correction.

B. NH listeners

Figures 2.9 and 2.10 display intelligibility for the individual NH listeners. Results for the SSN conditions are displayed in Fig. 2.9, and those for the babble conditions are displayed in Fig. 2.10. Note that the NH listeners were tested at the same SNRs as the HI listeners in SSN, but overlapped at only one SNR in babble. As anticipated, the performance of the NH listeners exceeded that of the HI listeners in unprocessed conditions. The mean NH scores for unprocessed stimuli were 64% and 86% correct for the two SSN SNRs (-5 and -2 dB) and 57% and 82% correct for the two babble SNRs (also -5 and -2 dB). Accordingly, algorithm benefit was considerably smaller for the NH than for the HI listeners. But some benefit was observed in 20 of the 24 cases for SSN and in all 24 of the cases for babble. Planned comparisons consisting of two-tailed paired *t*-tests on RAUs between unprocessed and processed scores for the NH listeners in each of the four conditions of Figs. 2.9 and 2.10 revealed significant benefit [each $t(11) \geq 3.9$, each two-tailed p value < 0.01]. This set of significant results also survives Bonferroni correction.

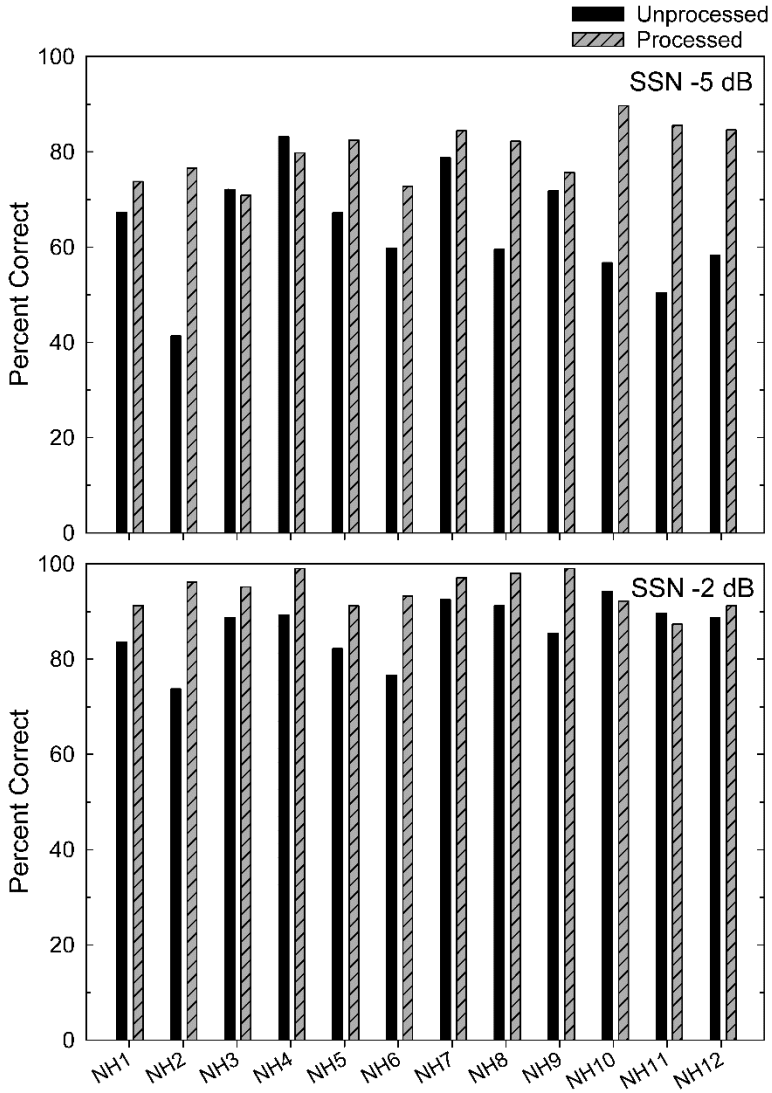

Figure 2.9. As Fig. 2.7, but for individual NH listeners and speech in SSN.

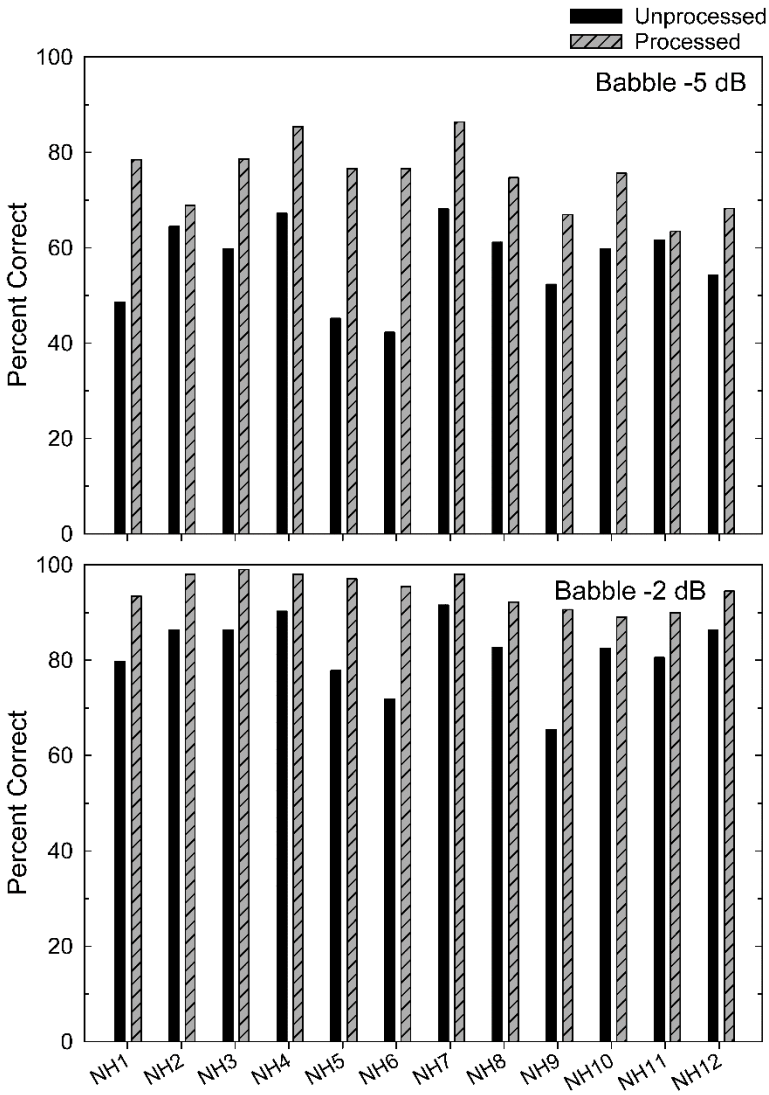

Figure 2.10. As Fig. 2.7, but for individual NH listeners and speech in multitalker babble.

C. Comparison between HI and NH listeners

Figure 2.11 displays group-mean sentence intelligibility scores and standard errors of the mean (SEMs) for both HI and NH listeners, plotted separately, in each condition. Again, SNRs are plotted in separate panels, black columns represent unprocessed scores, and shaded/hatched represent processed scores. The group-mean

algorithm benefit for the HI listeners was 46 and 48 percentage points for SSN at -5 and -2 dB SNR, respectively, and 58 and 53 percentage points for babble at -2 and 0 dB SNR, respectively. When benefit was expressed in RAUs to control for ceiling and floor effects, these values increased slightly to become 50 units for both SSN SNRs, and 60 and 57 units in babble. The figure also shows that the manipulation of SNR yielded the desired baseline (unprocessed) scores for the HI listeners. The mean baseline intelligibilities were 13% and 36% correct for SSN and 24% to 37% correct for babble. For the NH listeners, group-mean benefit values were 16 and 8 percentage points at -5 and -2 dB SNR, respectively, in SSN, and 18 and 13 percentage points at these SNRs in babble.

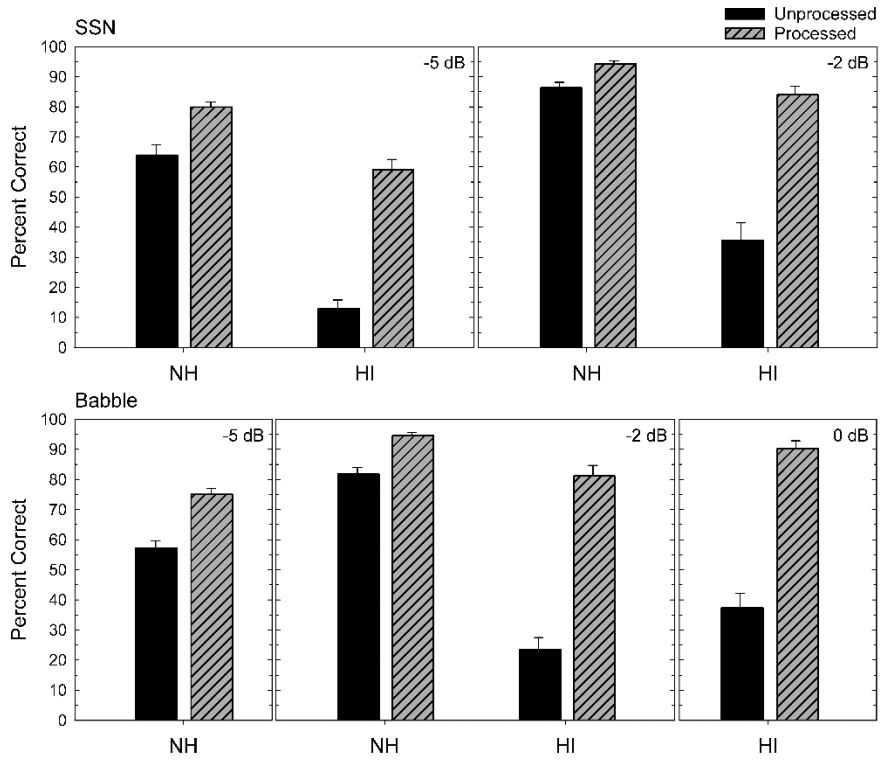

Figure 2.11. Group mean (and standard error) sentence intelligibility for the HI and NH listener groups. The background noise type and SNR are indicated, and the HI and NH groups are plotted separately. As in Figs. 2.7 – 2.10, benefit is reflected as the difference between a shaded/hatched column and the black column immediately to its left.

To address the question of whether the algorithm can restore NH speech-in-noise recognition abilities to these HI listeners, the performance of the HI listeners following algorithm processing was compared to the performance of young NH listeners without processing, in the conditions common to both groups. As Fig. 2.11 shows, the HI listeners approached within 1 percentage point of the NH listeners' performance in one condition (babble at -2 dB SNR) and within 5 percentage points in the remaining two conditions (SSN at -5 and -2 dB SNR). Three planned comparisons (two-tailed Welch's independent-samples *t*-tests on RAUs) between the algorithm-processed scores for the HI listeners and the unprocessed scores for the NH listeners in the three common conditions

indicated that differences were not significant [each $t \leq 1.1$, each $p > 0.3$, dfs adjusted using the Welch-Satterthwaite method ranged 17 - 22].

D. Comparison to Healy *et al*., 2013

To examine the human-subjects performance differences associated with upgrading to the current modern network architecture while substantially increasing the demands placed on the algorithm, the present results were compared with those of Healy *et al*. (2013). The present study was identical to Healy *et al*. (2013) in terms of the speech recordings used, the noise types and (most) SNRs employed, as well as the populations from which subjects were drawn, the numbers of subjects, the testing procedures, and the inclusion criteria for each listener group. The particular noises differed but were of the same type (SSN and babble). The primary difference across studies was the algorithm used for processing and the demands that had to be met.

In 2013, the algorithm was both trained and tested using the same talker and 10-s noise segments and it operated on future time frames, meaning it was neither talker, noise, nor corpus independent, nor was it causal. The present algorithm was required to generalize to an untrained talker from an untrained corpus and to untrained noise types as well as be fully causal. Figures 2.12 and 2.13 display group-mean sentence intelligibility scores and SEMs for HI and NH listeners in the conditions common to both studies, with SSN in the upper panels and babble in the lower panels of each figure. Pairs of columns labeled “2022” represent the current results and are replotted from Fig. 2.11, whereas pairs of columns labeled “2013” are from Healy *et al*. (2013). As with the previous

figures, benefit is represented as the difference between each unprocessed (solid column) and the corresponding processed score (hatched column). It is noted that baseline (unprocessed) scores differed across studies, likely attributable to the different samples of listeners employed and the use of different noises of the same type (different SSNs and different babbles).

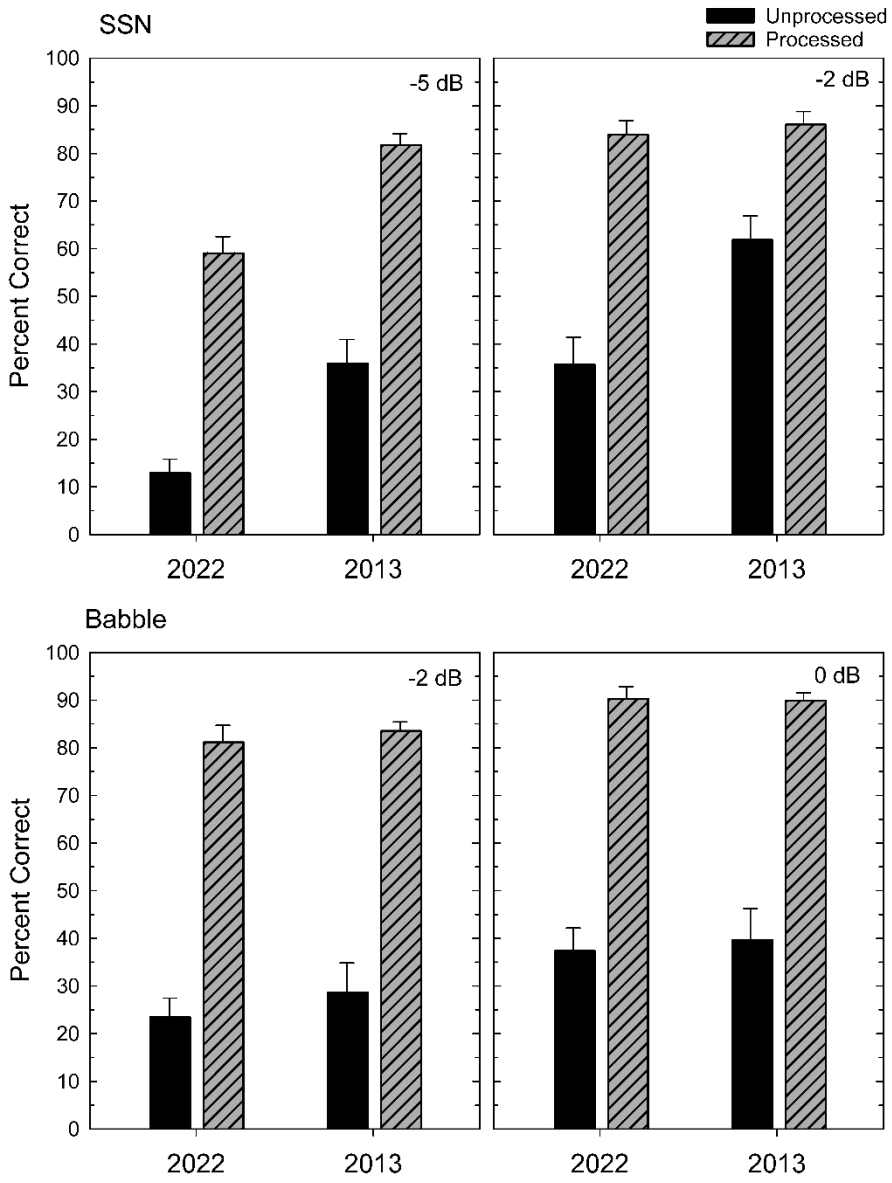

Figure 2.12. Comparison between the current study (2022) and the initial study (Healy *et al*., 2013). Shown are sentence intelligibility scores (and standard errors) for HI listeners, in two noise types, at different SNRs, both before and after algorithm processing. The comparison is between benefit obtained currently versus that obtained in 2013. Benefit is reflected as the height difference between each shaded/hatched column and the adjacent black column. The speech recordings, noise types, SNRs, testing procedures, and subject populations were identical across studies. The primary difference was the demand placed on the algorithm and the network architecture.

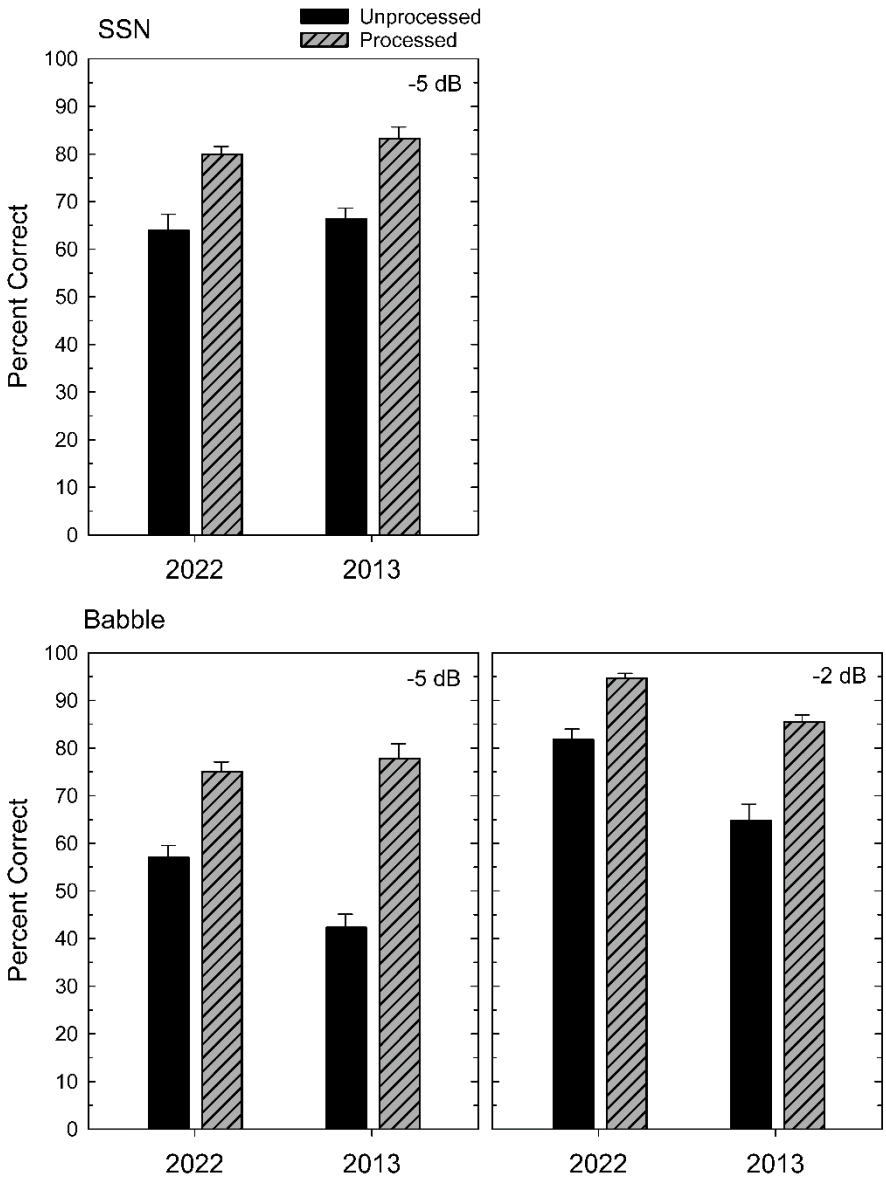

Figure 2.13. As Fig. 2.12, but for the NH listeners.

Figure 2.12 displays scores for the HI listeners. In SSN at -5 dB SNR (top left panel of Fig. 2.12), the group-mean algorithm benefit was similar across studies at approximately 46 percentage points, with only slightly higher benefit in the current study. However, group-mean algorithm benefit was considerably higher currently relative to

2013 (48 vs. 24 percentage points) at -2 dB SSN (top right panel of Fig. 2.12). For HI listeners in babble, mean benefit was slightly higher in the current study compared to 2013 at both -2 dB SNR (58 vs. 55 percentage points, bottom left panel of Fig. 2.12) and 0 dB SNR (53 vs. 50 percentage points, bottom right panel of Fig. 2.12).

Figure 2.13 displays scores for the NH listeners. In SSN, the SNR common across studies was -5 dB. In this condition, group-mean benefit was slightly lower in the current study compared to 2013 (16 vs. 17 percentage points, top panel of Fig. 2.13). Normal-hearing benefit was more noticeably lower for the current study than for 2013 in babble at -5 dB SNR (18 vs. 35 percentage points, bottom left panel of Fig. 2.13) and at -2 dB SNR (13 vs. 21 percentage points, bottom right panel of Fig. 2.13).

Planned comparisons consisting of two-tailed Welch's *t*-tests on RAUs were used to assess differences in group-mean algorithm benefit between the two studies. For HI subjects, mean benefit was numerically higher for the present (2022) algorithm in all 4 conditions, despite the far greater demands placed on it, but this difference was only significant in SSN at -2 dB [$t(18.6) = 3.1$, $p < 0.01$].

For NH subjects, there was no significant difference in benefit between the 2022 vs. 2013 algorithms in two out of the three conditions common to both studies (SSN at -5 dB SNR and babble at -2 dB SNR). In babble at -5 dB SNR, benefit was significantly higher for the 2013 algorithm [$t(20.1) = 3.68$, $p = 0.0015$].

E. Objective measures of intelligibility and sound quality

Table I displays objective measures obtained from acoustic analyses of the current stimuli. Scores for both noisy mixtures and processed mixtures are provided. Short-time objective intelligibility (STOI, Taal *et al*., 2011), represents a correlation between the amplitude envelope of the clean speech and that of the same speech once extracted from the mixture. Extended short-time objective intelligibility (ESTOI; Jensen and Taal, 2016), is similar and designed to better handle fluctuating noisy signals. Perceptual evaluation of speech quality (PESQ; Rix *et al*., 2001) is an objective measure of sound quality and ranges from -0.5 to 4.5. Finally, the scale invariant signal-to-noise ratio (SI-SNR, Le Roux *et al*., 2018) is an SNR estimate of the noisy and processed signals. STOI increased from noisy to ARN processed by 20 points, when averaged across conditions. ESTOI increased by an average of 32 points. PESQ increased by an average of 0.8, and SI-SNR increased by an average of 10 dB.

Table 2.1. Improvement in objective scores for sentences in two noises in different SNR conditions.

		STOI (%)		ESTOI (%)		PESQ		SI-SNR (dB)	
		Noisy	ARN	Noisy	ARN	Noisy	ARN	Noisy	ARN
Babble	-2 dB	62.6	84.9	38.4	72.6	1.47	2.37	-2.0	8.1
	0 dB	68.4	88.4	45.5	78.1	1.59	2.56	0.0	9.6
SSN	-5 dB	59.0	77.9	32.7	62.9	1.37	2.01	-5.0	6.2
	-2 dB	67.2	85.3	43.0	73.4	1.54	2.33	-2.0	8.7

IV. GENERAL DISCUSSION

The current results demonstrate that state-of-the-art deep learning noise reduction can produce large intelligibility improvements for HI and NH listeners, despite the considerable demands associated with extensive generalization and fully causal operation. Intelligibility benefit resulting from processing averaged 46 to 58 percentage points across conditions for the HI listeners. Benefit was lower for NH listeners (8 to 18 points), who typically experience less benefit. These benefits were statistically significant in all conditions.

The other goal of the current study was to compare the ability of the current algorithm to improve intelligibility relative to performance of the initial demonstration (Healy *et al*., 2013). This comparison provides a cumulative assessment of algorithm performance, or possible performance loss, resulting from the removal over time of various real-world constraints including talker, corpus, and noise dependence, as well as from causal operation. For HI listeners, it was found that the benefit observed currently matched or exceeded that observed in the initial study, but only significantly exceeded in one condition. For NH listeners, benefit was lower currently than in the initial study, but only significantly so in one condition.

It is concluded that modern deep learning based noise reduction can produce large intelligibility benefit for HI listeners (and also some benefit for NH listeners), despite the removal of multiple constraints and the resultant demanding test conditions. No decrement in benefit was observed for the HI listeners (but some decrement was observed for the NH listeners) resulting from the current algorithm relative to the initial

demonstration, thus illustrating the advancements made to neural network design and deep learning based noise reduction since 2013.

It was noted that baseline intelligibility in unprocessed conditions differed somewhat across the current study and the initial demonstration (see Figs. 2.12 and 2.13). However, all scores are free of strong floor and ceiling effects, and so they largely fall in the linear portion of the psychometric function relating intelligibility to information content. This allows benefit (differences across two points on this largely linear function) to be compared with reasonable confidence. Nevertheless, the comparison of exact benefit values across studies should not be emphasized.

Finally, it is noted that the current network remains large in size. Although computational complexity and the demand that a neural network places on the device on which it runs is not a fundamental aspect of viability, it is nonetheless an important consideration. Fortunately, the emerging field of model compression provides techniques that can be used to reduce the size of the model while retaining high performance.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and other Communication Disorders (R01 DC015521 to E.W.H., R01 DC012048 to D.L.W., and F32 DC019314 to E.M.J.) We gratefully acknowledge computing resources from the Ohio Supercomputer Center.

Notes

1. "Noise reduction" is used as an umbrella term to describe the isolation of target speech from various types of interference including background non-speech noise, speech babble, interfering speech from a single talker, room reverberation, and concurrent interferences involving more than one such interference. More specific terms for these processes include speech enhancement, speaker separation, and de-reverberation. "Single microphone" refers to conditions in which target speech and noise are received by the same single microphone and represents one of the most challenging but broadly applicable noise-reduction techniques.

2. Although there might be some value in having an algorithm operate optimally on the voice of a frequent communication partner (e.g., Tye-Murray *et al*., 2016).

3. See www.sound-ideas.com (Last viewed 5/21/2021).

REFERENCES

ANSI (**2004**). S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (American National Standards Institute, New York).

ANSI (**2010**). S3.6, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).

Ba, J. L., Kiros, J. R., and Hinton, G. E. **(2016).** "Layer normalization," arXiv:1607.06450.

Byrne, D., Parkinson, A., and Newall, P. (**1990**). "Hearing aid gain and frequency response requirements for the severely/profoundly hearing impaired," Ear Hear. **11**, 40–49.

Chen, J., Wang, Y., and Wang, D. L. **(2015)**. "Noise perturbation improves supervised speech separation," in *Proceedings of LVA/ICA*, pp. 83–90.

Chen, J., Wang, Y., Yoho, S. E., Wang, D. L., and Healy, E. W. (**2016**). "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," J. Acoust. Soc. Am. **139**, 2604–2612.

Goehring, T., Bolner, F., Monaghan, J. J. M., van Dijk, B., Zarowski, A., and Bleeck, S. (**2017**). "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," Hear. Res. **344**, 183–194.

Goehring, T., Keshavarzi, M., Carlyon, R. P., and Moore, B. C. J. (**2019**). "Using recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants," J. Acoust. Soc. Am. **146**, 705–718.

He, K., Zhang, X., Ren, S., and Sun, J. **(2016)**. "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Healy, E. W., Delfarah, M., Johnson, E. M., and Wang, D. L. (**2019**). "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," J. Acoust. Soc. Am. **145**, 1378–1388.

Healy, E. W., Delfarah, M., Vasko, J. L., Carter, B. L., and Wang, D. L. (**2017**). "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," J. Acoust. Soc. Am. **141**, 4230–4239.

Healy, E.W., Johnson, E.M., Delfarah, M., Sevich, V.A., Krishnagiri, D.S. and Wang, D. L. **(2021a)**. "Deep learning based speaker separation and dereverberation can generalize across different languages to improve intelligibility," J. Acoust. Soc. Am., **150**, 2526-2538.

Healy, E. W., Johnson, E. M., Delfarah, M., and Wang, D. L. (**2020**). "A talker-independent deep learning algorithm to increase intelligibility for hearing-impaired listeners in reverberant competing talker conditions," J. Acoust. Soc. Am. **147**, 4106-4118.

Healy, E. W., Tan, K., Johnson, E. M., and Wang, D. L. **(2021b)**. "An effectively causal deep learning algorithm to increase intelligibility in untrained noises for hearing-impaired listeners," J. Acoust. Soc. Am., **149**, 3943-3953.

Healy, E. W., Taherian, H., Johnson, E. M., and Wang, D. L. **(2021c)**. "A causal and talker-independent speaker-separation/dereverberation deep learning algorithm:

Cost associated with conversion to real-time capable operation," J. Acoust. Soc. Am., **150**, 3976-3986.

Healy E. W., Yoho S. E., Chen J., Wang Y., and Wang D. L. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**, 1660-1669.

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. L. **(2013)**. "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am.,. **134**, 3029-3038.

Hendrycks, D., and Gimpel, K. **(2016)**. "Gaussian error linear units (gelus)," arXiv preprint arXiv:1606.08415.

Hinton, G. E., Osindero, S., and Teh, Y.-W. **(2006)**. "A fast learning algorithm for deep belief nets," Neural Comput. **18**, 1527–1554.

Hochreiter, S., and Schmidhuber, J. **(1997)**. "Long short-term memory," Neural Comput. **9**, 1735–1780.

Jensen, J., and Taal, C. H. (**2016**). "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE/ACM Trans. Audio Speech Lang. Proc. **24**, 2009–2022.

Kearns, J. **(2014)**. "LibriVox: Free public domain audiobooks," Reference Reviews.

Keshavarzi, M., Goehring, T., Turner, R. E., and Moore, B. C. J. **(2019)**. "Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: A deep recurrent neural network and spectral subtraction," J. Acoust. Soc. Am. **145**, 1493–1503.

Kingma, D. P., and Ba, J. **(2014)**. "Adam: A method for stochastic optimization,: arXiv:1412.6980.

Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J.R. **(2018)**. "SDR – half-baked or well done?" arXiv:1811.02508v1.

Merity, S. **(2019)**. "Single headed attention RNN: Stop thinking with your head," arXiv:1911.11423.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., ... and Wu, H. **(2017)**. "Mixed precision training," arXiv:1710.03740.

Monaghan, J. J. M., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C. M., and Bleeck, S. (**2017**). "Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners," J. Acoust. Soc. Am. **141**, 1985–1998.

Nilsson, M., Soli, S. D., and Sullivan, J. A. **(1994)**. "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Panayotov, V., Chen, G., D., and Khudanpur, S. **(2015)**. "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210.

Pandey, A., and Wang, D. L. **(2020a)**. "On cross-corpus generalization of deep learning based speech enhancement," IEEE/ACM Trans. Audio. Speech Lang. Process. **28**, 2489–2499.

Pandey, A., and Wang, D. L. **(2020b)**. "Learning complex spectral mapping for speech enhancement with improved cross-corpus generalization," In INTERSPEECH, pp. 4511–4515.

Pandey, A., and Wang, D. L. **(2022)**. "Self-attending RNN for speech enhancement to improve cross-corpus generalization," IEEE/ACM Trans. Audio. Speech Lang. Process. **30**, 1374–1385.

Paul, D. B., and Baker, J. **(1992)**. "The design for the Wall Street Journal-based CSR corpus," In Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (**2001**). "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 749–752.

Santurette, S., Ng, E.H.N., Jensen, J.J., and Loong, B.M.K. (**2020**). "Oticon More clinical evidence," Oticon Whitepaper.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech, Lang., Hear. Res. **28**, 455–462.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Trans. Audio. Speech. Lang. Process. **19**, 2125–2136.

Tye-Murray, N., Spehar, B., Sommers, M., and Barcroft, J. (**2016**). "Auditory training with frequent communication partners," J. Speech, Lang., Hear., Res. **59**, 871–875.

Varga, A., and Steeneken, H. J. **(1993)**. "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Comm., **12**, 247-251.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. **(2017)**. "Attention is all you need," Advances in neural information processing systems, 30.

Wang, Y., and Wang, D. L. **(2013)**. "Towards scaling up classification-based speech separation," IEEE Trans. Audio. Speech Lang. Process. **21**, 1381–1390.

Wang, Y., Han, K., and Wang, D. L. **(2013)**. "Exploring monaural features for classification-based speech segregation," IEEE Trans. Audio. Speech Lang. Process. **21**, 270–279.

Zhao, Y., Wang, D. L. L., Johnson, E. M., and Healy, E. W. (**2018**). "A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions," J. Acoust. Soc. Am. **144**, 1627–1637.

Chapter 3. Manuscript 2

The optimal speech-to-background ratio for balancing speech intelligibility with environmental sound recognition

Eric M. Johnson and Eric W. Healy

Department of Speech and Hearing Science and Center for Cognitive and Brian Sciences

The Ohio State University, Columbus, OH 43210

Abstract

Speech often occurs in the presence of other sounds. The speech is usually regarded as the “signal,” but this is not always the case. At times, environmental sounds may be the signal of interest and speech may the “masker,” or both may simultaneously be of interest. Two experiments were carried out to determine the optimal speech-to-background ratio for jointly maximizing both speech intelligibility and environmental sound recognition. Normal-hearing (NH) and hearing-impaired (HI) listeners were tested in conditions of divided and/or selective attention. It was found that both NH and HI listeners are capable of high speech intelligibility and high environmental sound recognition over a range of speech-to-background ratios. The range and location of optimal speech-to-background ratios differed across NH and HI listeners. The optimal speech-to-background ratio also depended on the type of environmental sound present. These results have the potential to guide future noise-reduction designs that allow the listener access to highly intelligible speech without depriving them of a rich acoustic environment and access to potentially important environmental sounds.

I. INTRODUCTION

Listening often occurs in the presence of multiple sound sources. Under these conditions, the acoustic wave that arrives at each ear is a mixture of sounds. When sounds originate from different directions relative to the listener's head, binaural cues can aid in perceptually segregating a signal of interest from the other sound(s) in the auditory scene (Bregman, 1990). However, in what may be described as a "worst-case scenario" where sounds arrive from the same direction, the listener must rely primarily on monaural cues to segregate sounds from one another. For humans, the speech of a target talker is often a signal of interest. Because other sounds can reduce speech intelligibility via masking, they are traditionally considered to be the "noise," "interference," "masker," or "background."

An extensive body of research has examined speech intelligibility in the presence of a wide variety of masker types, including sine waves, square waves, and repetitive pulses (Stevens *et al*., 1946); warbling tones, music, bandpass filtered noise, amplitude-modulated noise, frequency-modulated noise, peak-clipped noise, gated noise, and one or more competing talkers (Miller, 1947); white noise (Hawkins and Stevens, 1950; Fletcher, 1953); complex tones (Licklider and Guttman, 1957); speech-shaped noise (Busch and Eldredge, 1967); speech-modulated speech-shaped noise (Brungart, 2001); spectro-temporally modulated noise (Howard-Jones and Rosen, 1993; Fogerty *et al*, 2018); environmental sounds such as naval ship noises (Klumpp and Webster, 1963) and aircraft noise (Kryter and Williams, 1966); and complex soundscapes such as cafeteria noise (Cooper and Cutts, 1971) and traffic noise (Aniansson, 1978). Generally, speech

intelligibility for listeners with normal hearing (NH) is unaffected by broadband noise as long as the signal-to-noise ratio (SNR) is at least 6 dB, and performance can remain above chance even when the SNR is as low as -18 dB (Licklider and Miller, 1951). The effect of noise on speech intelligibility for hearing-impaired (HI) listeners tends to be greater, and the magnitude of this effect largely depends on their degree, type, and configuration of hearing loss (Ross *et al*., 1965).

However, the speech of a single target talker may not always be the only signal of interest in an acoustic mixture. In many situations, there are other sounds that are also relevant to the listener, such as environmental sounds that provide information about objects and events around them. Environmental sounds not only contribute to listeners' sense of awareness and well-being (Jenkins, 1985; Ramsdell, 1978; Reed and Delhorne, 2005) but also warn them of potential dangers in the environment (e.g., approaching vehicles, sirens, falling objects).

When a relevant environmental sound overlaps with target speech, each is both signal and masker because each is of interest to the listener, and each has the potential to interfere with the perception of the other. Because of this, the term "speech-to-background ratio" (SBR) will be used in place of "signal-to-noise ratio" in the present study to express the amplitude relationship between speech and a concurrent environmental sound.

Whereas many studies, including several of those cited above, have examined speech intelligibility in the presence of competing noise, few if any have considered environmental sound recognition (ESR) in the presence of competing speech. The present

study is concerned with both of these aspects of auditory perception and the conditions under which both are possible. It is known that speech intelligibility improves toward ceiling performance as SBR increases, and it presumed that, inversely, ESR decreases toward the performance floor with increasing SBR. There thus exists a trade-off between speech intelligibility and ESR as SBR varies. In the context of this trade-off, the present study aims to answer the following primary questions: (i) Is it possible to jointly obtain high speech intelligibility and high ESR? (ii) What is the optimal average SBR for jointly maximizing both speech intelligibility and ESR for NH listeners? (iii) Does this value differ for a group of bilateral hearing aid users with HI? (iv) Is there a range of SBRs over which speech intelligibility and ESR are both high? (v) Is this range different for NH and HI listeners? And (vi) does the optimal SBR value depend on the type of background sound? The answers to these questions will further our understanding of auditory perception in normal and impaired systems. Moreover, since environmental sound recognition is an important priority for adults with hearing loss (Bell, 2005), the knowledge gained in this study could inform the design of future hearing devices, with the aim of presenting signals at the optimal SBR for providing both high speech intelligibility and allowing for high ESR.

A secondary research question involves the costs associated with performing two tasks simultaneously (divided attention). There are often limitations on listeners' abilities to process all available auditory information in parallel, and divided attention concerns the optimal allocation of cognitive resources between different signals by splitting or rapidly shifting attentional focus (Parasuraman, 1998). Numerous studies have examined

the human performance costs associated with divided attention when both tasks are speech recognition (e.g., Humes *et al*., 2006; Gallun *et al*., 2007; Fumero et. al, 2022). The present study aims to establish whether divided attention results in similar performance costs when one of the tasks is ESR, which involves a different set of auditory processing mechanisms (Lewis *et al*., 2004).

In Experiment 1, subjects performed speech recognition and ESR concurrently (divided attention). In Experiment 2, they performed each task separately (selective attention). Both experiments were concerned with listeners' auditory perception in the absence of binaural cues, thus creating a worst-case scenario in terms of cues available to separate speech and environmental sound. Psychometric functions for each task in each condition were generated to address the research questions listed above.

II. Experiment 1

In Experiment 1, speech intelligibility and ESR were assessed in conditions of divided attention: Listeners were instructed to attend to both the speech and the environmental sound on each trial. This represents a dual-task paradigm that is more challenging than the many everyday listening situations where listeners selectively attend to only one sound source at a time. It therefore also represents a lower bound for human performance for these particular tasks.

A. METHOD

1. Subjects

Two groups of listeners participated in Experiment 1. The first group was composed of 11 adults with NH, defined as pure-tone thresholds at 20 dB HL or lower at octave frequencies from 250 to 8000 Hz (ANSI, 2004, 2010). The two exceptions were a listener with a threshold of 25 dB HL at 1000 Hz in one ear and another listener with a threshold of 30 dB HL at 8000Hz in one ear. The NH subjects were recruited from undergraduate courses at The Ohio State University and received course credit for participating. Ages ranged from 18 to 20 years (mean = 20), and all were female. Young listeners with NH were selected for the current task to represent an upper bound for human performance.

The second group consisted of 10 HI listeners who were recruited from The Ohio State University Speech-Language-Hearing Clinic and selected to represent a diverse sample of adults with hearing aids. Accordingly, they were all binaural hearing aid users with varying degrees of bilateral sensorineural hearing loss. Hearing status was confirmed on day of test using pure-tone audiometry (ANSI, 2004, 2010). Pure-tone-average thresholds (PTAs), based on audiometric thresholds at 500, 1000, and 2000 Hz and averaged across ears, ranged from 13 to 59 dB hearing level (HL) with a mean of 30. Eight of the HI listeners had at least one audiometric threshold within normal limits (20 dB HL or lower) in at least one ear, but all of them had moderate to profound hearing loss (thresholds greater than 40 dB HL) at two or more frequencies in both ears, or thresholds in the impaired range (greater than 20 dB HL) at half or more of the audiometric frequencies in both ears. Hearing-impaired subjects were numbered in order of increasing PTA (i.e., higher subject numbers correspond to greater mid-frequency hearing loss).

Each audiogram displayed in Fig. 3.1 is labeled with an HI listener's subject number, age, and sex. Ages ranged from 27 to 67 years (mean = 51), with six females and four males. HI subjects received a monetary incentive for participating.

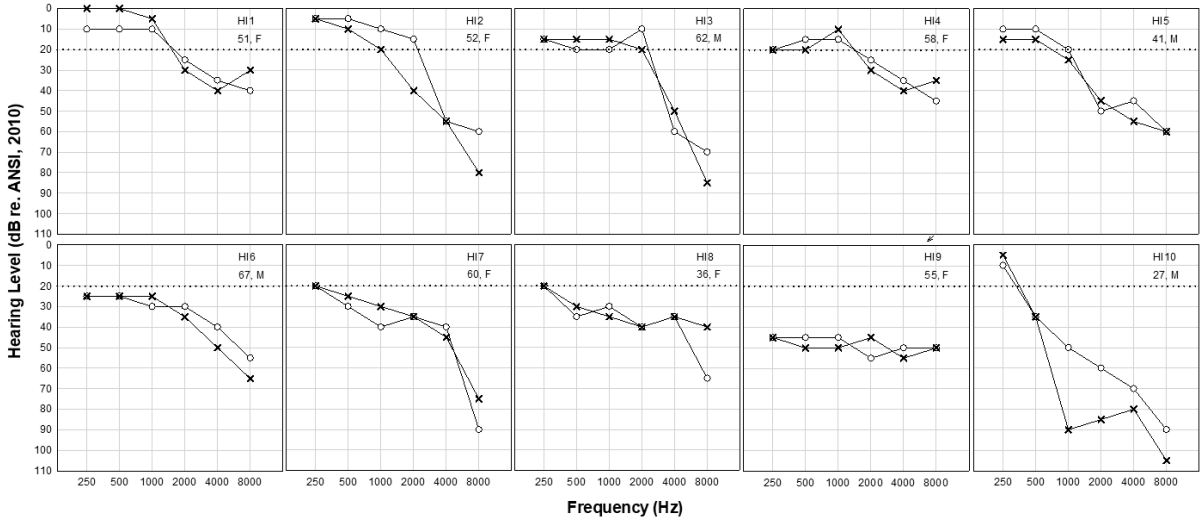

Figure 3.1. Pure-tone air-conduction audiometric thresholds for the listeners with hearing impairment. Listeners are numbered in order of increasing degree of hearing loss. Right ears are represented by circles and left ears are represented by X's. The limit of normal hearing (20 dB HL) is represented by a dotted horizonal line in each panel. Subject numbers, ages in years, and sexes are also provided.

All participants spoke American English as their native language, and none had any previous experience with the sentences used in this study. All participants gave their informed written consent to participate in this research, which was approved by the Institutional Review Board (IRB) at The Ohio State University.

2. Stimuli

Each stimulus consisted of a sentence mixed with an environmental sound. Sentences were drawn from the standard recording of the Hearing in Noise Test (Nilsson *et al*., 1994) and were all produced by an adult male native speaker of General American English.

Environmental sounds were sourced from the Database of Environmental Sounds for Research Activities (Gygi and Shafiro, 2010), the Sound Effects Recognition Test (Finitzo-Hieber *et al.*, 1980), a commercial stock media website (Pond5.com), and field recordings. The 25 sounds selected for this study were judged to be familiar to the general population and sufficiently distinct from each other to minimize confusion between sounds. Importantly, the duration of each included sound was at least as long as the longest sentence stimulus used in the study. This ensured that sounds were capable of masking entire sentences via simultaneous masking. Thus, transients and other short-duration sounds were not used as environmental sounds. See Figure 3.2 for the 25 environmental sounds used.

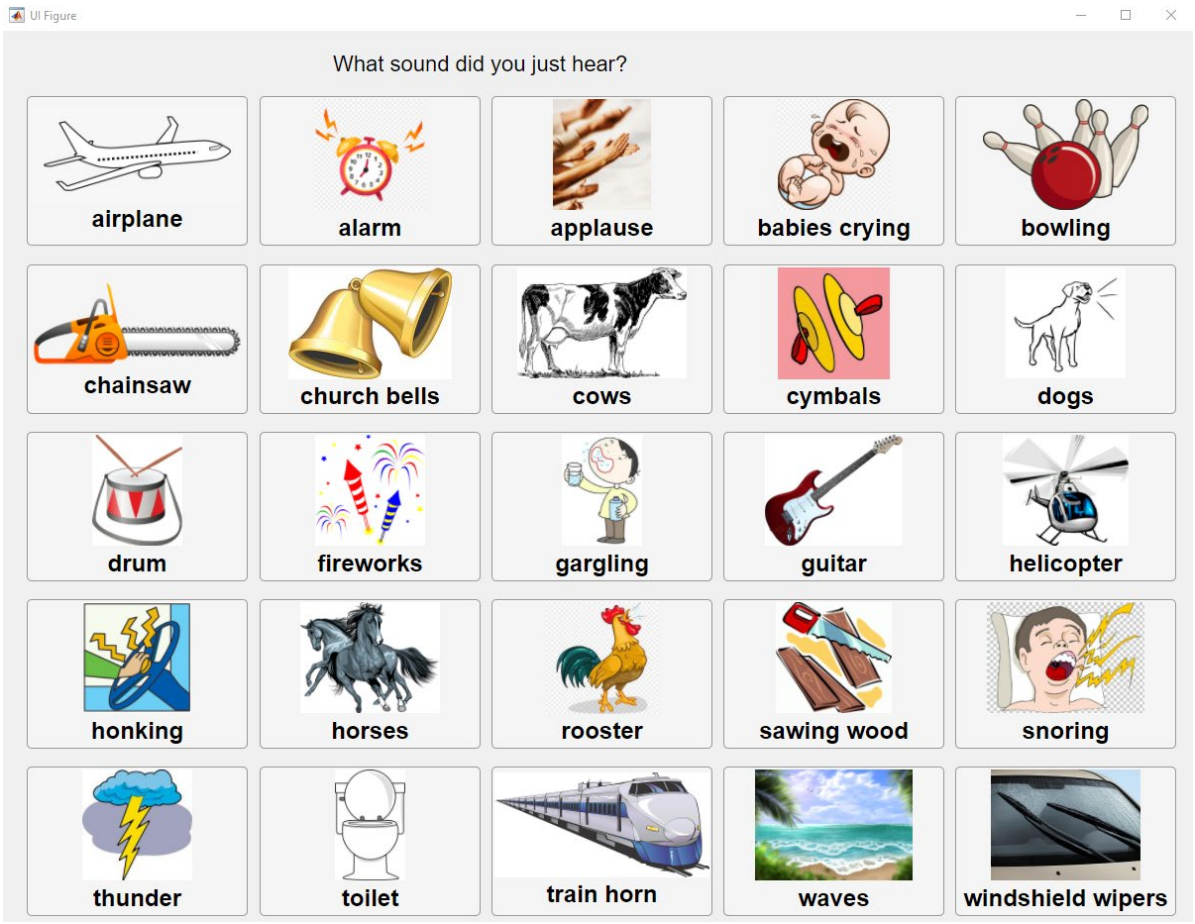

Figure 3.2. The graphical user interface that listeners used to record their responses for environmental sounds. The 25 environmental sounds are arranged in a 5 x 5 grid in alphabetical order. Pictures are provided to assist in locating the appropriate response button.

The environmental sound to be mixed with the sentence was chosen randomly for each stimulus. Stimuli were mixed at the following 10 SBRs: -36, -30, -8, -4, 4, 20, 32, 40, 64, and 70 dB. These SBRs were selected based on pilot testing to produce the entire range of ceiling to floor scores for both speech intelligibility and ESR, with at least two points on the steep portion and two points on the lower asymptote of each psychometric function.

Prior to mixing, silence was trimmed from the beginning and end of each sentence and environmental sound so that their onsets would align, thus eliminating any precursor

or preceding fringe containing only speech or only background. Next, the length of the environmental sound was trimmed to match the length of the sentence, thus aligning their offsets and eliminating glimpses containing only the background sound at the end each stimulus. The environmental sound was then rescaled to achieve the desired SBR and mixed with the sentence. Finally, each mixture was scaled to the same root mean square (RMS) amplitude. The sampling rate and bit depth of all stimuli were 44.1 kHz and 16 bits, respectively.

3. Procedure

Each subject heard a total of 250 stimuli in a single session, blocked by the 10 SBR conditions, with 25 stimuli per condition. The SBRs were presented in a random order for each subject, whereas the sentences were presented in the same fixed order for each subject. Again, the background sound was chosen randomly, with replacement, on each experimental trial. Thus, there were always 25 possible background sounds for each trial, and a given sound could be used repeatedly or not at all in a condition. But on average across listeners, each environmental sound was presented once per listener per SBR.

Individual listeners were tested in a double-walled audiometric booth, seated in front of a computer monitor and mouse. The experimenter was also seated in the booth, approximately six feet from the listener. Listeners were instructed to perform two tasks on each trial containing a sentence mixed with an environmental sound: First, repeat the sentence out loud; then, click on the labeled picture corresponding to the background

sound in the stimulus (Fig. 3.2), guessing on either task if unsure. Listeners were instructed to perform the tasks in this order to ensure they fully completed their spoken response before the onset of the subsequent stimulus presentation, which occurred 300 ms after the environmental-sound response. The experiment was thus self-paced. The experimenter recorded the number of words correctly reported in each sentence while the custom MATLAB application recorded listeners' responses to the ESR task. Both listener and experimenter were blind to the conditions under test. No feedback was provided to listeners during formal testing.

Stimuli were played back on a Windows PC, converted to analog form using an RME Fireface UCX audio interface (Haimhausen, Germany), amplified using a Mackie 1202-VLZ mixer (Woodinville, WA), and presented diotically over Sennheiser HD 280 Pro headphones (Wedemark, Germany). Hearing-impaired listeners were tested without their hearing aids. The level of each stimulus was set to 65 dBA in each ear for the NH subjects and verified using a sound-level meter and flat-plate coupler (Larson Davis models 824 and AEC 101, Depew, NH). For the HI listeners, additional individualized frequency-specific gains, as prescribed by the NAL-RP formula (Byrne *et al*., 1990), were applied to this 65 dBA presentation level, to compensate for their hearing loss. These gains were implemented using a RANE DEQ 60L digital equalizer (Mukilteo, WA), as described in Healy *et al*. (2015). The sound pressure level following NAL-RP amplification did not exceed 96 dBA for any participant.

Prior to the start of formal testing, each participant completed a three-stage familiarization. The first stage involved learning to recognize the 25 background sounds

in the absence of any competing speech. Participants heard each of the 25 background sounds, presented in a random order, and were instructed to click on the labeled picture representing the sound. Feedback was provided for any incorrect responses during the first stage of familiarization. This process was repeated until the listener could identify each sound without prompting from the experimenter. One potential subject who could not perform this task was dismissed from the experiment. This stage was repeated up to 3 times, with an average across listeners of 2 repetitions required. In the second stage of familiarization, listeners heard seven HINT sentences in quiet and repeated each one out loud. In the third familiarization stage, listeners heard seven sentences mixed with a random background sound at each of the following three SBRs, in this order: 12, -8, and 40 dB. Listeners were instructed to repeat back as much of each sentence as possible and then click on the picture representing the background sound, taking their best guess if unsure. The 28 HINT sentences used for familiarization were distinct from the 250 used for testing. Following the three familiarization stages, formal testing began, as described above. In total, each experimental session lasted approximately one hour.

B. RESULTS

Psychometric functions based on the cumulative normal function were fit for each task type and listener group using the *quickpsy* package (Linares and López-Moliner, 2016) in R 4.1.3 (R Core Team, 2022). In each function, the explanatory variable was SBR, and the response variable was either words or environmental sounds correctly reported. The lapse rate (i.e., the probability of an incorrect response, independent of

stimulus intensity) was calculated as a free parameter. The guess rate was set to 0 for the speech intelligibility functions and to 0.04 for the ESR functions, representing chance performance for a closed-set task with 25 response alternatives. The range of each function was then normalized such that lower and upper asymptotes were mapped to 0 and 100% correct, in order to account for guessing and lapses and thus facilitate comparisons between open-set and closed-set tasks.

Figure 3.3 displays normalized psychometric functions for speech intelligibility and ESR based on the pooled performance of the 11 NH listeners. Filled circles represent normalized percent words correct for the speech-intelligibility task whereas open circles represent normalized percent-correct ESR. The solid black line is the fitted psychometric function for intelligibility, and the dashed line is the fitted function for ESR.

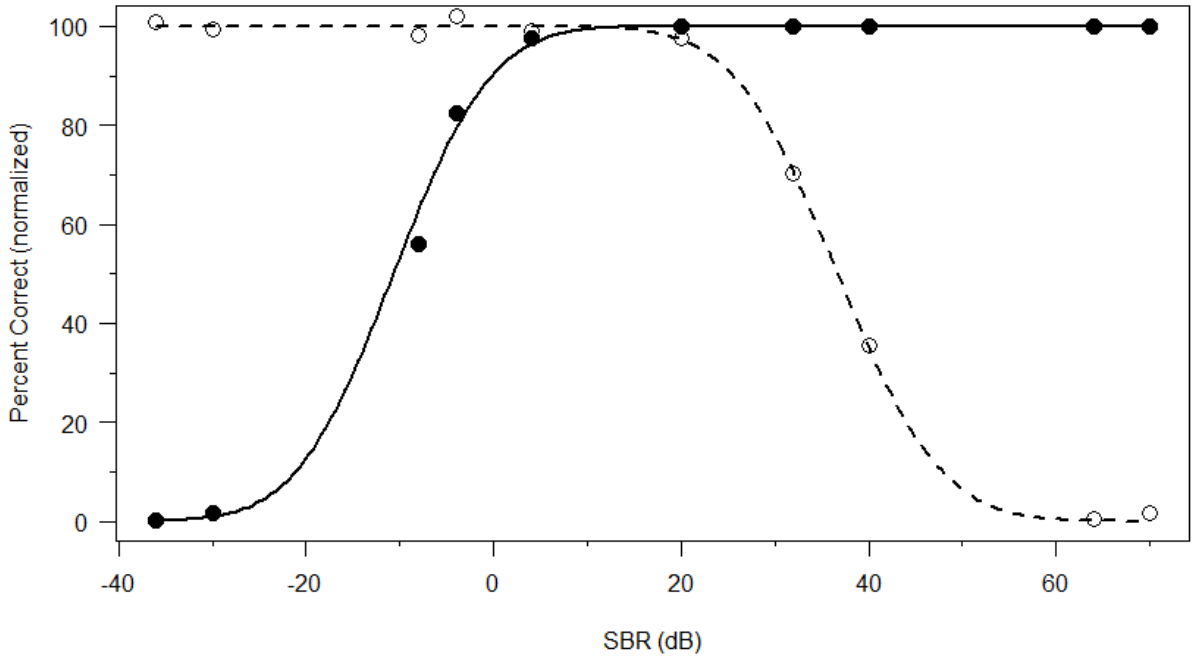

Figure 3.3. Normalized psychometric functions for speech intelligibility and environmental sound recognition based on the pooled performance of 11 normal-hearing listeners. Filled circles denote percent words correct speech intelligibility and open circles correspond to percent correct environmental sound recognition. The solid black line is the fitted psychometric function for intelligibility, and the dashed line is the fitted function for environmental sound recognition.

The estimated lapse rate for NH speech intelligibility was 0.0005, 95% CI [0.00095, 0.0011], indicating that NH listeners accurately performed the speech-recognition task when conditions were favorable. Their threshold for 50% normalized intelligibility was -10.7 dB SBR, 95% CI [-11.0, -10.2], and their threshold for 95% normalized intelligibility was 2.8 dB SBR, 95% CI [2.26, 3.32].

The estimated lapse rate for NH ESR was 0.0274, 95% CI [0.0198, 0.0345], indicating that maximum unnormalized performance plateaued at approximately 97.6% correct. The NH listeners' normalized threshold for 50% correct ESR was 36.7 dB SBR, 95% CI [37.7, 35.8]. Their normalized threshold for 95% correct ESR was 22.4 dB SBR, 95% CI [24.7, 19.4].

Thus, the range between NH listeners' normalized 50% correct thresholds for speech intelligibility and ESR was 47.4 dB (-10.7 to 36.7 dB SBR), whereas the range between their 95% correct thresholds spanned 19.6 dB (2.8 to 22.4 dB SBR). Since performance for both tasks was high (95% correct or greater) for all SBRs within these limits, any SBR bounded by 2.81 and 22.4 dB could be regarded as sufficiently optimal to essentially maximize both speech intelligibility and ESR. However, a single optimal value, based on the data collected in the present study, may be calculated as the point of intersection between the two psychometric functions. This value is 12.2 dB SBR, where predicted normalized performance for both speech intelligibility and ESR equal 99.7% correct.

Figure 3.4 displays normalized psychometric functions for the 10 HI listeners who participated in this study. As Fig. 3.3, mean SBR for both speech intelligibility and ESR

are shown. Filled black circles represent normalized percent words correct, and open circles represent normalized percent environmental sounds correct. Again, the solid black line is the psychometric function for speech intelligibility, and the dashed line is the fitted curve for ESR.

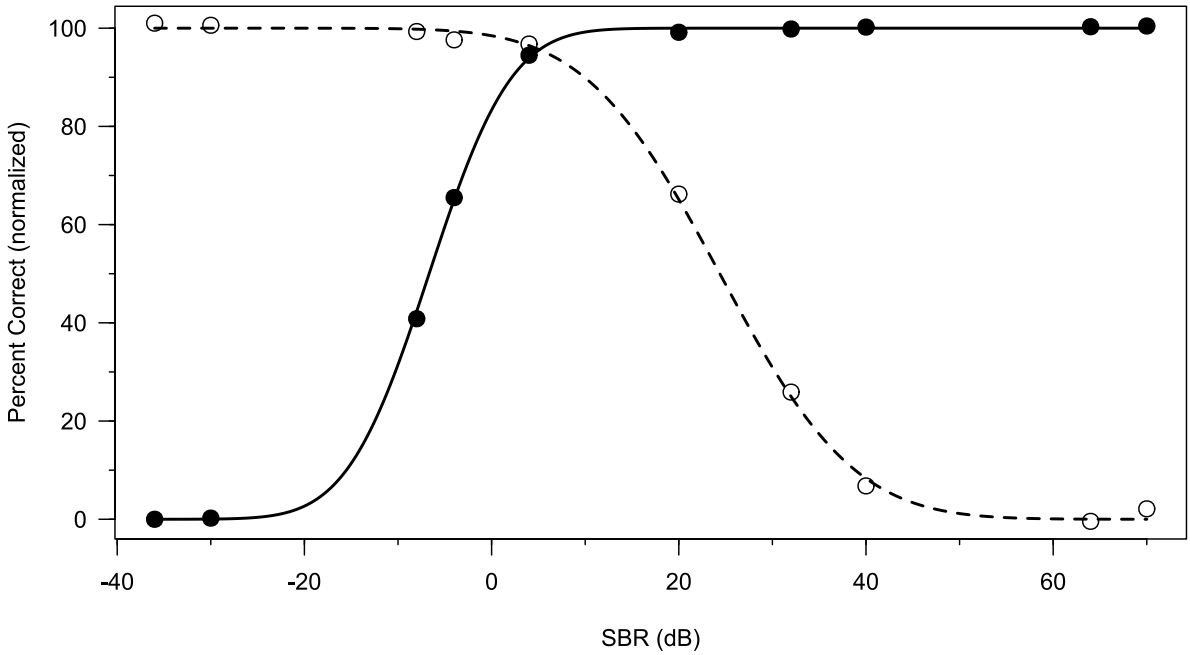

Figure 3.4. As Fig. 3.3, but for the hearing-impaired listeners

For the HI listeners, the estimated lapse rate for speech recognition was 0.005, 95% CI [0.0038, 0.0071], indicating that the HI listeners also achieved high performance on this task with few lapses under favorable conditions. The HI listeners' normalized 50% threshold for speech intelligibility was -6.7 dB SBR, 95% CI [-7.0, -6.3], and their normalized 95% threshold was 4.7 dB SBR, 95% CI [3.9, 5.4].

For ESR, the estimated HI lapse rate was 0.018, 95% CI [0.009, 0.028], which indicates, perhaps surprisingly, that the HI listeners were less prone to lapses on the ESR

task than the NH listeners were. Their normalized 50% and 95% thresholds for ESR were 24.2 (95% CI [25.5, 23.2]) and 5.8 dB SBR (95% CI [3.5, 8.6]), respectively.

The estimated range between normalized 50% thresholds for speech intelligibility (-6.7 dB SBR) and ESR (24.2 dB SBR) for these HI listeners was 30.9 dB. There was also a narrow 1.1-dB range between which the HI listeners' normalized thresholds for speech intelligibility and ESR were both above 95% correct. This range extended between 4.7 and 5.8 dB SBR. For these HI listeners, the optimal SBR for balancing the trade-off between speech intelligibility and ESR, when both are considered equally important, was calculated to be 5.1 dB, which is the point where the two psychometric functions intersect, at 95.6% normalized percent correct. It is further noted that other SBRs may be optimal when speech intelligibility is prioritized. For example, at 10 dB SBR, normalized speech intelligibility is above 99% correct while normalized ESR remains relatively high at 90% correct.

Compared to the NH listeners, the HI listeners' group threshold for 50% speech intelligibility was 4 dB higher, and their group threshold for 50% ESR was 12.5 dB lower (both poorer). Thus, it appears that the smaller range over which the HI listeners were able to recognize both speech and environmental sounds was more driven by an inward shift of the ESR function than by an inward shift of the psychometric function for speech intelligibility. As a group, HI and NH listeners also differed in terms of their optimal SBR for maximizing both intelligibility and ESR. This value was 5.1 dB SBR for the HI listeners and 12.2 dB SBR for the NH listeners. Although these particular values differ

for NH vs. HI listeners, the optimal values for both listener groups still allows normalized speech intelligibility and ESR to both be above 96% correct.

III. EXPERIMENT 2

In this experiment, listeners performed the speech recognition and ESR tasks separately rather than simultaneously. The primary purpose of this experiment was to compare the human performance of these tasks across a wide variety of environmental sound types. The secondary purpose was to compare human performance in conditions of selective vs. divided attention.

A. METHOD

1. Subjects

A group of 20 young adult listeners with NH participated in Experiment 2. They all had pure-tone thresholds of 20 dB HL or better at octave frequencies from 250 to 8000 Hz (ANSI, 2010), with the exception of one listener with a threshold of 25 dB HL at 250 Hz in one ear. These NH subjects were also recruited from and received credit in undergraduate courses at The Ohio State University. Their ages ranged from 18 to 26 years (mean = 19), and all were female. Young listeners with NH were selected for Experiment 2 because the goal was to establish the basic performance differences between divided and selective attention in the current task, without additional possible factors associated with aging and/or hearing loss. These listeners were distinct from those who participated in Experiment 1, and none of them had any prior exposure to the

sentence materials employed in this experiment. Again, and all participants gave informed consent to participate, and all procedures were approved by the IRB.

2. Stimuli

As in Experiment 1, stimuli for Experiment 2 consisted of sentences drawn from the standard recording of the HINT, each of which was mixed with an environmental sound that was randomly selected from the same set of 25 sounds used in Experiment 1. A set of 110 HINT sentences was used for the speech recognition task, and a separate set of 110 HINT sentences was used for the ESR task. For each task, stimuli were mixed at 11 SBRs. The SBRs used in the speech recognition task were -35, -30, -25, -20, -15, -10, -5, 0, 5, 10, and 15 dB. The SBRs used in the ESR task were 0, 7, 14, 21, 28, 35, 42, 49, 56, 63, and 70 dB. Other than the SBRs employed, the process for mixing and scaling the stimuli was the same as in Experiment 1.

3. Procedure

The same general procedures that were used in Experiment 1 were followed in Experiment 2, unless otherwise noted. Listeners were testing in a single session consisting of an ESR task followed by a speech-recognition task.

a. Environmental sound recognition task

Listeners heard 330 stimuli in the ESR task, blocked by SBR, with 10 trials per block. For each of the 11 SBRs, there were three blocks of 10 stimuli. These 30 SBR

blocks were presented in a random order for each listener. The same set of 110 HINT sentences was repeated three times in the experiment, with sentences presented in the same order each time. The repetition of sentences was allowed because the task only involved ESR and not sentence recognition. Listeners were instructed to attend to the background sound, click on the corresponding labeled picture, and guess if unsure. As in Experiment 1, the playback sequence of each stimulus was triggered by the listener's response to the previous one.

Prior to formal testing, listeners were familiarized with the 25 environmental sounds, as described in Experiment 1. They also practiced identifying sounds in the presence of competing speech using the MATLAB application and 11 stimuli mixed at 6, 10, 14, 18, 22, 26, 40, 44, 48, 52, and 56 dB, with difficulty of the task increasing with each successive practice trial. The sentences used during familiarization were distinct from those used for formal testing. After completing all experimental trials for the ESR task, listeners next performed the speech-recognition portion of the experiment.

b. Speech-recognition task

For the speech-recognition task, listeners heard 110 stimuli and were instructed to attend to the talker's voice and repeat back each sentence as best as they could, guessing if unsure. The sentences were blocked by SBR, with 10 sentences per block. The experimenter controlled stimulus presentation and recorded the number of words correctly reported by the listener in each trial.

Prior to formal testing, listeners practiced recognizing speech in background noise using 11 sentences mixed at 16, 12, 8, 4, 0, -4, -8, -12, -16, -20, and -24 dB SBR, with speech recognition becoming more difficult with each successive practice trial. Again, the sentences used in familiarization were not used during formal testing. Experiment 2 lasted approximately one hour for each participant, including both tasks and familiarization.

B. Results

Figure 3.5 displays the normalized group-mean percent-correct word identification scores for the speech recognition task (filled black circles) and the group-mean percent-correct sound identification scores for the separate ESR task (open circles) at each SBR tested. Psychometric functions for each task type were again fitted using the same linking function, software, and other parameters outlined in Experiment 1. The solid black and dashed lines are the normalized psychometric functions for speech intelligibility and ESR, respectively.

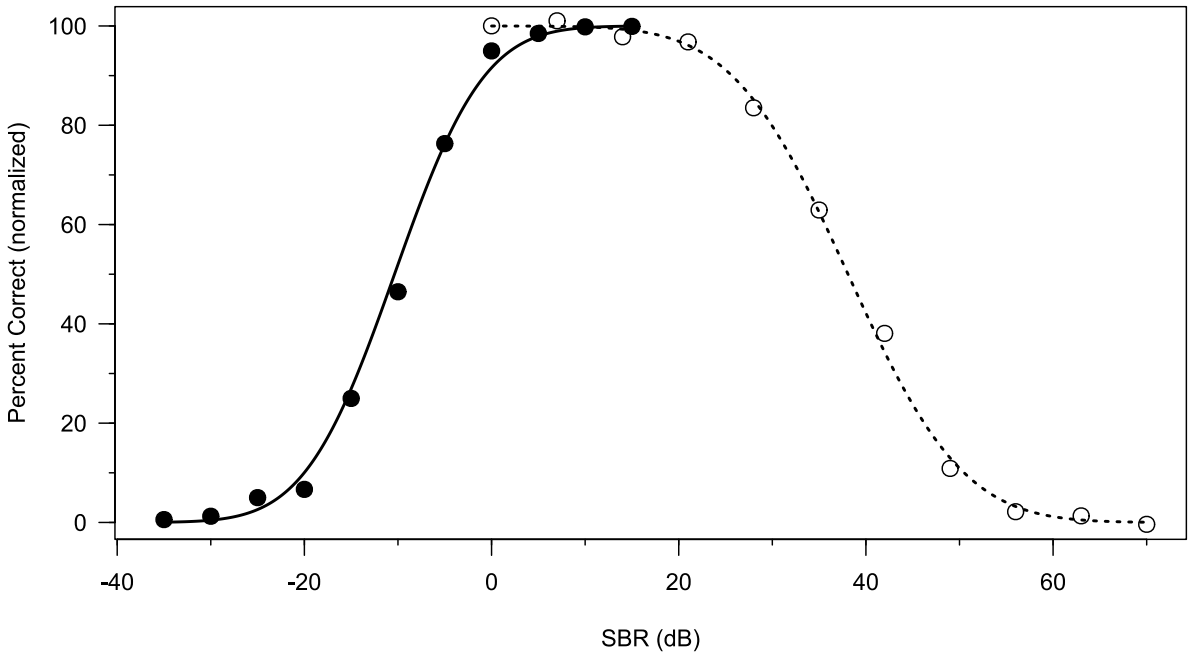

Figure 3.5. Normalized psychometric functions for speech intelligibility (drawn in a solid black line) and environmental sound recognition (drawn in a dashed line) for 20 normal-hearing subjects. The filled circles denote mean normalized speech intelligibility at 11 signal-to-background ratios, and the open circles denote mean normalized environmental sound recognition scores at a different set of 11 signal-to-babble ratios.

For speech intelligibility, performance was near perfect when the SBR was sufficiently high, with an estimated lapse rate of less than 0.0001. Combined with a guess rate of 0, normalization had a negligible effect on the psychometric function for intelligibility. The normalized threshold for 50% HINT words correct was -10.4 dB SBR, 95% CI [-10.6, -10.1], and the normalized 95% threshold was 2.0 dB SBR, 95% CI [1.5, 2.5].

The lapse rate for ESR was estimated to be 0.027, 95% CI [0.02, 0.035], indicating that unnormalized performance plateaued at approximately 97.3% correct. The normalized 50% and 95% correct thresholds were 38.1 (95% CI, [38.5, 37.7]) and 22.2 (95% CI, [23.3, 21.3]) dB SBR, respectively.

The estimated distances between the normalized 50% and 95% correct thresholds for speech intelligibility and ESR were 48.5 and 24.2 dB, respectively. The optimal SBR for maximizing both intelligibility and ESR, defined as the intersection between the two normalized psychometric functions, was 10.8 dB, at which point normalized performance was 99.8% correct for both tasks.

Figure 3.6 displays psychometric functions for speech intelligibility and ESR based on subsets of the entire dataset for the 25 environmental sound stimuli used in this study, with one panel per environmental sound. The repetition of sentences in Experiment 2 produced a number of trials sufficient to allow this analysis of individual environmental sounds. Panels are arranged in order of decreasing range between normalized 50% correct thresholds for speech intelligibility and ESR. Filled black circles represent percent-correct speech intelligibility scores, and solid black curves represent the corresponding fitted psychometric functions. Percent-correct scores and fitted curves for ESR are indicated by open circles and dashed lines, respectively. For each environmental sound and task type, the lapse rate was modeled as a free parameter unless accuracy was 100% at the three highest SBRs for intelligibility or the three lowest SBRs for ESR, in which case the lapse rate was set to 0. The lapse rate was also set to 0 if a negative value was calculated. The guess rate was set to 0 for speech intelligibility and to 0.04 for ESR. All other relevant fitting procedures and parameters are described in Experiment 1. For each environmental sound, if the two normalized psychometric functions crossed at a single point, the optimal SBR for maximizing both speech intelligibility and ESR was calculated as the point of intersection. If both psychometric functions plateaued at 100%

correct (normalized) and overlapped over a range of SBRs, then the optimal SBR value was calculated as the midpoint between normalized 50% correct thresholds.

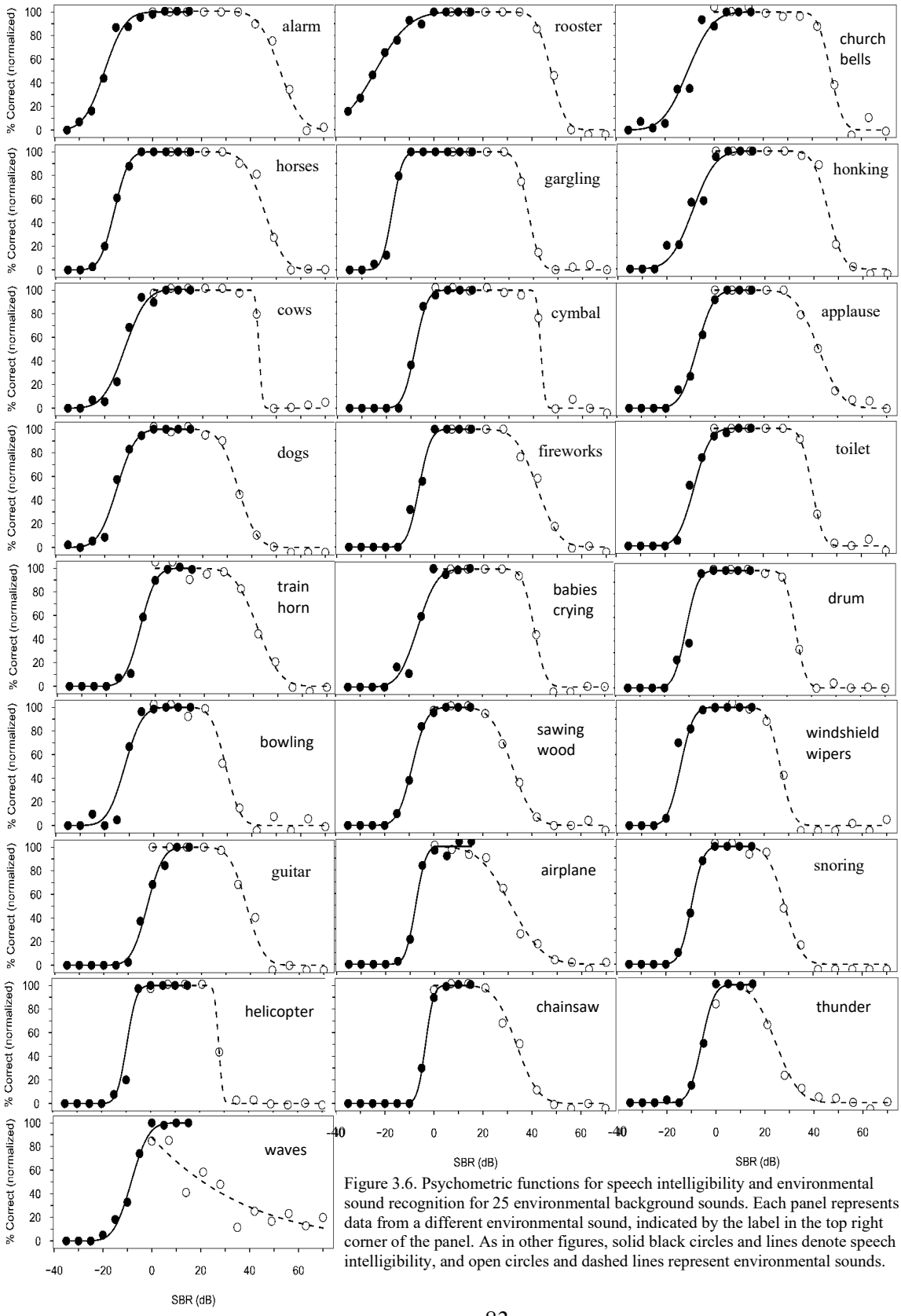

Figure 3.6. Psychometric functions for speech intelligibility and environmental sound recognition for 25 environmental background sounds. Each panel represents data from a different environmental sound, indicated by the label in the top right corner of the panel. As in other figures, solid black circles and lines denote speech intelligibility, and open circles and dashed lines represent environmental sounds.

Apparent in Fig. 3.6 is that all environmental sounds could be reliably identified, with the exception of "waves." It appears that listeners tended to guess "waves" by default when they were unable to hear the target environmental sound. Whereas many other sounds were unlikely to be guessed at high SBRs, as indicated by ESR scores below the chance rate of 4% correct (or below 0% correct after normalization), ESR scores for "waves" were well above chance, even at the highest SBR of 70 dB. Furthermore, ESR for "waves" never reached ceiling performance, even at the lowest SBR tested for selective attention ESR, which was 0 dB.

Table 1 contains a confusion matrix for selective-attention ESR. Stimulus sounds are listed vertically in the first column whereas subject responses are listed horizontally in the first row of the table, both in alphabetical order. Each cell indicates the number of responses given to each stimulus type across SBRs and listeners. The total number of presentations of each stimulus and listener responses of each stimulus are given in the last column and row, respectively. There were between 217 and 297 total presentations of each stimulus type (M = 264, SD = 17.2). Even though the number of presentations for "waves" (277) was near the mean, the number of "waves" responses (805) was 4.4 standard deviations above the mean number of responses (264), indicating a listener bias for this response.

Table 3.1. Confusion matrix for environmental sound recognition in the presence of competing speech and under conditions of selective attention. Stimuli presented are listed in the first column, and listener response types are shown in the first row. The number in each cell represents the number of listener responses given to each environmental sound stimulus type.

stimulus \ response	airplane	alarm	applause	babies crying	bowling	chainsaw	churchbells	cows	cymbals	dogs	drum	fireworks	gargling	guitar	helicopter	honking	horses	rooster	sawing wood	snoring	thunder	toilet	train horn	waves	windshield wipers	Total
airplane	121	2	13	1		5	5	4	2	1	6	9	2	5	8	4	8	4	3		6	4	3	32	5	253
alarm	1	181	3	1	1	1	2		4	4	1	6	4	2	4	1	1	1	7	1	4	4	1	21		256
applause	2	1	157		2	1	7	3	2	1	6	5	3	2	1	3	2	2	5	2	10	4	4	36	4	265
babie scrying	6	2	5	162	2	5	1	5	2	1	3	4	4	8	3	1	4	2	7	1	5	3	2	21	2	261
bowling	2	3	6		118	1	4	6	3		5	12	2	8	3	3	3	2	14	3	11	6	4	34	7	260
chainsaw		2	8	1	3	120	4	4	4		6	3	3	11	3	2	6	5	18	1	9	8	1	28	8	258
churchbells			6	1	1		180	8	2	2		6	7		1	1	4	3	5	1	12	2	3	18	4	267
cows	3	2	5	1		2	2	156	1	1	4	3	1	6	3	3	7	5	12		1	3	2	24	7	254
cymbals	3	1	4	1	2	3	4	4	167	1	7	5	2		3	3		7	7	1	3	4	3	33	6	274
dogs	3		4		10	2	6	6	3	139	5	4	7	2	7	2	6	7	13	2	3	5		24	8	268
drum	1	2	6	1	4	1	3	3	5	4	136	10	1	9	3	5		7	11		8	5	7	30	4	266
fireworks	4	2	7		3	2	1	2	5	2	6	160	3	5	2	2	6	4	8	1	5	3	2	29	2	266
gargling	2	1	5		3	6	2	2	6	4	2	7	155	4	5	2	6	4	9		6			24	3	258
guitar	5	1	8	2	3	8	3		3	3	4	4	4	167	3	3	4	2	12	2	7	7	5	34	3	297
helicopter	2	3	14	2	4	3	3	6	1	4	3	7	7	2	107	2	4	4	11	3	9	10	3	50	5	269
honking		1	7	3	2			4	7	2	5	6	5	4	4	182		1	4		6	5	4	25	8	285
horses	2	1	1		2	3	1	7	6	2	7	2	5	3	2	5	184	9	2	1	11	4	3	20	3	286
rooster	3		2	2	5	2	1	1	2	1	4	5	5	1	2		5	185	2		6	4	4	13	3	258
sawing wood	2	1	8	1	5	4	4	6	5	6	2	3	4	3	4	6	17	4	137	3	8	6		33	4	276
snoring	3	1	9	1	3	5	4	2	7	2	9	8	5	4	3	4	5	4	19	97	11	4	1	37	6	254
thunder	5	2	10		8	4	2	5	1	3	4	19	5	5	8		6		8	3	90	13	3	53	11	268
toilet	1	2	3		1	5	4	3	6	2	7	4	5	2	4			2	7		9	138	3	16	1	225
train horn	3	1	5	3	3	4	2	1	2	1	3	5	2	3	2	27	8	6	3		5	2	156	31	4	282
waves	7		7	1	5	3	5	2	2	4	6	5	4	6	12	4	4	5	9	5	44	9	2	114	12	277
windshield wipers	2	2	8	2	4	5	6	2		1	6	5	7	4	4	1	4	4	20	1	7	7	1	25	89	217
Total	**183**	**214**	**311**	**186**	**194**	**195**	**256**	**242**	**248**	**191**	**247**	**307**	**252**	**266**	**201**	**266**	**294**	**279**	**353**	**128**	**296**	**260**	**217**	**805**	**209**	**6600**

Table 2 lists optimal SBR values, normalized 50% correct thresholds for speech intelligibility and ESR, and widths of ranges between 50% thresholds for each of the 25 environmental sounds. Sounds are listed in descending order of distance between normalized 50% correct thresholds. Thresholds for normalized 50% correct speech intelligibility ranged from -23.8 dB SBR for "rooster" to -2.2 dB SBR for "guitar" dB SBR, with a mean of -10.4 (*SD* = 5.0). Thresholds for normalized 50% correct ESR ranged from 21.0 dB SBR for "waves" to 52.2 dB SBR for "alarm" with a mean of 37.2 (*SD* = 8.2). Thresholds for intelligibility and ESR appeared to be weakly inversely related, but the Pearson's correlation coefficient was not significant [$r(23) = -0.33$, $p =$

0.11]. The width of the range between normalized 50% correct thresholds was as small as 29.4 dB for “waves” and as high as 71.8 dB for “alarm.” Optimal SBRs for balancing intelligibility with ESR for different environmental sounds also varied, ranging from -1.2 to 18 dB SBR for “waves” and “horses,” respectively (M = 12.0, SD = 5.5), indicating that the optimal SBR value depends on the type of environmental sound present.

Table 3.2. Optimal SBR values, normalized 50% correct thresholds for speech intelligibility and ESR, and widths of ranges between 50% thresholds for each of the 25 environmental sounds.

Stimulus Name	Optimal SBR (dB)	Normalized 50% correct threshold for speech intelligibility (dB SBR)	Normalized 50% correct threshold for environmental sound recognition (dB SBR)	Width of range between 50% thresholds (dB)
alarm	16.3	-19.6	52.2	71.8
rooster	12.0	-23.8	47.7	71.5
church bells	14.5	-16.0	45.1	61.2
horses	18.3	-10.8	47.4	58.1
gargling	10.1	-17.6	37.8	55.4
honking	18.2	-9.4	45.7	55.1
cows	15.5	-11.9	42.9	54.8
cymbals	17.3	-8.3	43.0	51.3
applause	17.4	-7.4	42.2	49.6
dogs	9.7	-15.0	34.5	49.5
fireworks	17.8	-6.4	41.9	48.3
toilet	15.8	-8.2	39.8	48.0
train horn	13.3	-6.1	41.7	47.7
babies crying	17.3	-6.5	41.1	47.6
drum	11.1	-11.2	33.4	44.7
bowling	8.8	-11.8	29.4	41.2
sawing wood	11.6	-8.8	32.0	40.8
windshield wipers	6.5	-14.1	26.6	40.7
guitar	18.1	-2.2	38.5	40.7
airplane	1.5	-7.7	31.1	38.9
snoring	9.4	-9.6	28.4	37.9
helicopter	9.0	-9.7	27.8	37.5
chainsaw	6.7	-3.4	33.7	37.1
thunder	4.8	-5.9	24.6	30.4
waves	-1.2	-8.4	21.0	29.4

The secondary research question for Experiment 2 involves the effect of selective vs. divided attention for NH listeners performing the mixed tasks of speech intelligibility and ESR. The overall normalized 50% correct threshold for selective-attention speech

intelligibility was -10.4 dB SBR, compared to -10.7 dB SBR for divided attention, a difference of 0.3 dB. For ESR, the overall normalized thresholds for 50% correct were 38.1 and 36.7 dB for selective and divided attention, respectively, amounting to a difference of 1.4 dB.

To test the effect of selective vs. divided attention on speech intelligibility and ESR, two mixed-effects logistic models were constructed in R 4.1.3 (R Core Team, 2022) using the lme4 (Bates *et al.*, 2014) and lmerTest (Kuznetsova *et al.*, 2017) packages. In the first model, the outcome variable was binary data for words correct (n = 26,109 observations), and the fixed effects were SBR, attention type (selective vs. divided), and their interaction. SBR was expressed in dB units that were otherwise untransformed, and attention type was deviation coded, with selective attention coded as 0.5 and divided attention coded as -0.5. The model included random slopes for listener. Models for speech intelligibility that included random slopes for sentence and/or and environmental sound failed to converge.

In the second model, the outcome variable was binary data for environmental sounds correct (n = 9,350 observations). This model also had fixed effects for SBR (in dB), attention type, and their interaction. As in the first model, attention type was deviation coded. The model for ESR included random slopes for sentence, listener, and environmental sound. Models that included random slopes failed to converge.

Unsurprisingly, the effect of SBR was large and significant in both models. Increases in SBR reliably predicted higher speech intelligibility (β = 0.22, SE = 0.003, z = 64.6, $p < 0.0001$) and lower ESR performance (β = -0.14, SE = 0.003, z = -44.5, $p <$

0.0001). The effect of attention type was significant for both intelligibility (β = -0.58, *SE* = 0.16, z = 3.73, p = 0.0002) and ESR (β = 2.0, *SE* = 0.23, z = 8.64, $p < 0.0001$) with selective attention resulting in higher performance on both tasks, compared to divided attention. The interaction between SBR and attention type was also significant in both models. For speech intelligibility, the effect of selective attention was greater at higher SBRs (β = 0.061, *SE* = 0.007, z = 8.91, $p < 0.0001$) whereas the effect of selective attention was greater at lower SBRs for ESR (β = -0.052, *SE* = 0.005, z = -10.2, $p <$ 0.0001).

IV. DISCUSSION

The results of these experiments illustrate the trade-off between speech intelligibility and ESR when SBR is varied. They also demonstrate that it is possible to find a balance between these two competing aims by selecting an SBR where performance is high for both speech recognition and ESR. Both NH and HI listeners are capable of high speech intelligibility and high ESR over a range of SBRs, and this range is larger for NH listeners. It is also possible to identify a single SBR that, on average across a variety of environmental sounds, yields the highest combined performance of speech intelligibility and ESR. This optimal SBR value is not the same for NH and HI listeners, but the optimal SBR for HI listeners still yields very high performance for NH listeners. Furthermore, the optimal NH SBR also yields very high speech intelligibility (above 99% correct) for HI listeners while still providing for good ESR (approximately 86% correct).

Different sounds interact differently with speech. For some sounds, such as the alarm used in this study, the range over which intelligibility and ESR are both high is large. This may be a particularly useful quality of an alarm sound: it is easily audible and identifiable in the presence of a competing speech signal, and yet it does not unduly disrupt the transmission of speech information.

The present study also demonstrated the effect of attention on speech intelligibility and ESR. The two types of attention tested in this study, fully divided and fully selective, represent two theoretical endpoints. Thus, these results show the largest drop in performance that is likely to arise from the division of attention when the two tasks are speech recognition and ESR. The costs of divided attention resulted in poorer performance on both the speech recognition and ESR tasks. Although reliable and highly significant in the statistical models, the drops in performance associated with divided attention were small in terms of raw dB values (0.3 dB threshold shift for intelligibility and 1.4 dB threshold shift for ESR). Accordingly, the results obtained with regard to balancing speech intelligibility and ESR are not largely affected by the exact task that the listener is asked to perform.

The results of this study can also be used to inform the design of noise-reduction algorithms. Historically, noise reduction has simply not been very effective, and so much of the background remains following processing. Accordingly, the goal of noise-reduction systems has been to remove as much background as possible. However, advances coming largely from deep learning have allowed strict isolation of the target speech and suppression of the background. This new ability to isolate suggests that the

goal needs to be reconsidered. As the current work shows, it is possible to retain high speech intelligibility and environmental sound awareness using an appropriate target SBR value. It is suggested that the goal of future more highly effective noise reduction algorithms shift toward this goal.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and other Communication Disorders (R01 DC015521 and F32 DC019314).

REFERENCES

ANSI (**2004**). S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (American National Standards Institute, New York).

ANSI (**2010**). S3.6, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).

Aniansson, G. (**1978**). "Speech intelligibility in and speech interference levels of traffic noise in hearing-impaired and normal listeners," Acta Oto-Laryngologica, **86**, 109-112.

Apoux, F., Carter, B. L., & Healy, E. W. (**2018**). "Effect of dual-carrier processing on the intelligibility of concurrent vocoded sentences," J. Speech Lang. Hear. Res. **61**, 2804-2813.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (**2014**). "Fitting linear mixed-effects models using lme4," arXiv preprint arXiv:1406.5823.

Bell, B. The Psychological/Social Impact of Cochlear Implants [thesis]. Rochester, NY: Rochester Institute of Technology Scholar Works; **2005**;138. Available at: https://scholarworks.rit.edu/theses/8088/.

Bregman, A. S. (**1990**) *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101-1109.

Busch, A. C., & Eldredge, D. (**1967**). "The effect of differing noise spectra on the consistency of identification of consonants," Language and Speech, **10**, 194-202.

Cooper Jr, J. C., & Cutts, B. P. (**1971**). "Speech discrimination in noise," Journal of Speech and Hearing Research, **14**, 332-337.

Finitzo-Hieber, T., Gerling, I. J., Matkin, N. D., & Cherow-Skalka, E. (**1980**). "A sound effects recognition test for the pediatric audiological evaluation," Ear Hear. **1**, 271-276. https://doi.org/10.1097/00003446-198009000-00007

Fletcher, H. (**1953**). *Speech and Hearing in Communication*. New York: Van Nostrand (reprinted by the Acoustical Society of America, 1995).

Fogerty, D., Carter, B. L., & Healy, E. W. (**2018**). "Glimpsing speech in temporally and spectro-temporally modulated noise," J. Acoust. Soc. Am. **143**, 3047-3057.

Gygi, B., & Shafiro, V. (**2010**). "Development of the database for environmental sound research and application (DESRA): Design, functionality, and retrieval considerations," EURASIP J. Audio Speech Mus. Process., 1-12. https://doi.org/10.5281/zenodo.2622626

Hawkins Jr, J. E., & Stevens, S. S. (**1950**). "The masking of pure tones and of speech by white noise," J. Acoust. Soc. Am. **22**, 6-13.

Howard-Jones, P. A., & Rosen, S. (**1993**). "Uncomodulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. 93(5), 2915-2922.

Humes, L. E., Lee, J. H., & Coughlin, M. P. (**2006**). "Auditory measures of selective and divided attention in young and older adults using single-talker competition," J. Acoust. Soc. Am. **120**, 2926-2937.

Jenkins, J. J. (**1985**). "Acoustic information for objects, places, and events," In W. H. Warren, & R. E. Shaw (Eds.), Persistence and change, (pp. 115–138). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Klumpp, R. G., & Webster, J. C. (**1963**). "Physical Measurements of Equally Speech-Interfering Navy Noises," J. Acoust. Soc. Am. **35**, 1328-1338.

Kryter, K. D., & Williams, C. E. (**1966**). "Masking of speech by aircraft noise," J. Acoust. Soc. Am. **39**, 138-150.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (**2017**). "lmerTest package: tests in linear mixed effects models," Journal of statistical software, **82**, 1-26.

Lewis, J. W., Wightman, F. L., Brefczynski, J. A., Phinney, R. E., Binder, J. R., & DeYoe, E. A. (**2004**). "Human brain regions involved in recognizing environmental sounds," Cerebral cortex, **14**, 1008-1021.

Licklider, J. C. R. & Miller, G. A. (**1951**) "The perception of speech," In: Stevens SS (ed) Handbook of Experimental Psychology. New York: John Wiley, pp. 1040–1074.

Licklider, J. C. R., & Guttman, N. (**1957**). "Masking of speech by line-spectrum interference," J. Acoust. Soc. Am. **29**, 287-296.

Linares, D., & López-Moliner, J. (**2016**). "quickpsy: An R package to fit psychometric functions for multiple groups," The R J. **8**, 122-131.

Miller, G. A. (**1947**). "The masking of speech," Psych. Bull. **44**, 105-129.

Parasuraman R. (**1998**). "The attentive brain: Issues and prospects," In: Parasuraman R, editor. *The Attentive Brain.* Cambridge, Massachusetts: MIT Press, pp. 3–15.

R Core Team. (**2022**). “R: A language and environment for statistical computing,” R Foundation for Statistical Computing. http://www.R-project.org/

Ramsdell, D. A. (**1978**). “The psychology of the hard-of-hearing and the deafened adult,”. In H. David, & S. R. Silverman (Eds.), *Hearing and deafness*. New York: Holt, Rinehart & Winston.

Reed, C. M., & Delhorne, L. A. (**2005**). “Reception of environmental sounds through cochlear implants,” Ear Hear. **26**, 48-61.

Ross, M., Huntington, D. A., Hayes, A. N., & Dixon, R. F. (**1965**). “Speech discrimination of hearing-impaired individuals in noise,” J. of Aud. Res. **5**, 47–72.

Stevens, S. S., Miller, J., & Truscott, I. (**1946**). “The masking of speech by sine waves, square waves, and regular and modulated pulses,” J. Acoust. Soc. Am. **18**, 418-424.

Chapter 4: Manuscript 3

An ideal compressed mask for increasing intelligibility without eliminating environmental sound recognition

Eric M. Johnson, Eric W. Healy

Department of Speech and Hearing Science and Center for Cognitive and Brian Sciences

The Ohio State University, Columbus, OH 43210

Abstract

Hearing-impairment is often characterized by poor speech-in-noise recognition. Modern noise-reduction algorithms are capable of providing listeners with increased speech intelligibility, but when the goal of a system is to eliminate all portions of the signal that are not the target speech, this increased intelligibility comes at the expense of environmental sound recognition. Environmental sound recognition is an important part of the human auditory experience that not only provides a sense of connection to the environment but also forecasts potential safety hazards in the vicinity. This paper proposes a modified version of the ideal ratio mask, known as the ideal compressed mask, that aims to provide listeners with improved speech intelligibility without sacrificing environmental sound awareness. This is accomplished by limiting the maximum attenuation that the mask can perform on a time-frequency unit. In a dual-task paradigm, speech intelligibility and environmental sound recognition for hearing-impaired and normal-hearing listeners were measured using stimuli that had been processed by ideal compressed masks with various levels of maximum attenuation. It was found that this type of processing resulted in significantly improved intelligibility and high environmental sound recognition performance for both types of listeners. It was also found that the same level of maximum attenuation provided the optimal balance of intelligibility and environmental sound recognition for both listener types. It is argued that future deep learning based noise reduction algorithms may provide overall better

outcomes for listeners by targeting an ideal compressed mask rather than a time-frequency mask that seeks to eliminate all but the target speech from the signal.

I. INTRODUCTION

Speech perception in background noise can represent a very significant challenge for a variety of listeners. Although normal-hearing (NH) listeners can typically tolerate considerable amounts of noise if conditions are otherwise ideal, even these best listeners struggle to understand speech when the signal-to-noise ratio is sufficiently low (Licklider and Miller, 1951). For hearing-impaired (HI) listeners, this struggle can be much greater, and it represents one of their primary auditory complaints (see Moore, 2007; Dillon, 2012). In fact, poor speech recognition in noise negatively affects quality of life by contributing to social isolation, depression, dependence, frustration, loneliness, and communication difficulties (Ciorba *et al.*, 2012). Thus, there is a substantial need to solve this "speech-in-noise" problem, especially for HI listeners.

Fortunately, techniques now exist to address this problem by improving speech intelligibility in noise. One strategy that can be used to increase intelligibility is called time-frequency (T-F) masking, a process in which an acoustic mixture is divided in both time and frequency into small units that are selectively attenuated based on the signal-to-noise ratio (SNR) of each unit: units where target speech dominates are preserved, and units where noise dominates are attenuated. This technique can be used to isolate speech from noise, and it serves as one of the main building blocks for a number of effective deep learning based noise-reduction algorithms (Wang *et al.,* 2013; 2014; *Healy et al*., 2013, 2015, 2019; Chen *et al*., 2016; Zhao *et al*., 2018).

Time-frequency masking may take several forms, each of which can have its own advantages and disadvantages in terms of sound quality, benefit to intelligibility, and

suitability for machine learning based classification algorithms. The classical form of T-F masking is the ideal binary mask (IBM), which was originally proposed as a benchmark for measuring the segregation performance of computational auditory scene analysis systems (Hu and Wang, 2001; Wang, 2005). The IBM assigns each T-F unit a value of 1 if it is dominated by the target speech or 0 if it is dominated by noise. The IBM is then multiplied with the T-F representation of the speech-plus-noise mixture, causing units dominated by the target speech to remain intact and units dominated by the noise to be discarded.

The ideal ratio mask (IRM; Srinivasan *et al*., 2006; Narayanan and Wang, 2013; Hummerstone *et al*., 2014; Wang *et al*., 2014) is similar to the classic Wiener filter (see Loizou, 2007). Like the IBM, each T-F unit is assigned an attenuation value based on its SNR. But instead of limiting attenuation values to the binary options of 0 and 1, the IRM can assign any value along a continuum from 0 to 1, theoretically allowing for an infinite number of possible attenuation values. As with the IBM, the IRM is multiplied by the T-F array of the speech-plus-noise mixture to scale each T-F unit according to its speech versus noise dominance. Units with the lowest SNRs are attenuated the most while units having the highest SNRs are attenuated the least. Another type of T-F mask is known as the ideal quantized mask (IQM; Healy and Vasko, 2018), which may be considered a hybrid of the IBM and the IRM. Like the IBM, it scales T-F units in discrete steps, but like the IRM, it can assign more than only two possible attenuation values.

Each of these ideal T-F masks can substantially improve the intelligibility of noisy speech. Brungart *et al*. (2006), Li and Loizou (2008a,b), Kim *et al*. (2009), Kjems

et al. (2009), and Sinex (2013) all found that the IBM could yield near-perfect sentence intelligibility for NH listeners in various noise backgrounds. Anzalone *et al.* (2006) and Wang *et al*. (2009) found that the IBM could significantly improve NH and HI listeners' speech-reception thresholds for sentences in different noises. Regarding the IRM, Madhu *et al*. (2013) and Koning *et al*. (2015) found that it can also deliver near 100% correct sentence intelligibility for NH listeners in different noise backgrounds, including multi-talker and single-talker interference. The IQM has been shown to provide the intelligibility and sound-quality advantages of the IRM while retaining the classification-based nature of the IBM, which may have algorithmic advantages (Healy and Vasko, 2018).

Today's modern noise-reduction algorithms, many of which are based on some form of T-F masking, are designed to maximize speech intelligibility by removing as much interference from the signal as possible. And although these algorithms considerably improve speech understanding (see Manuscript I), they also limit access to environmental sounds, potentially "deafening" users to all nonspeech sounds. This creates a new hearing deficit even as it solves another. Similar to improved speech understanding, access to nonspeech environmental sounds is an important priority for adults with hearing loss (Bell, 2005). Environmental sound awareness allows for greater personal independence and an improved sense of "connection" with one's surroundings (Harris *et al*., 2017). The ability to perceive and respond to environmental sounds is also essential for personal safety and danger avoidance, and hearing loss is associated with an increased risk of workplace and nonworkplace injuries requiring medical care (Mick *et*

al., 2018). Many potential hazards are often forecast by environmental sounds, either naturally occurring (e.g., the sound of an approaching car) or artificial (e.g., alarms, alerts, and warning signals; U.S. Fire Administration, 1999). Other examples of safety-relevant sounds include gunshots, emergency sirens, vehicle horns, impact noises, and even the warning whistle of a lifeguard or crossing guard. The inability to detect and resulting failure to react to these environmental sounds could result in potentially fatal consequences. Critically, therefore, any noise reduction technology intended for hearing protheses must achieve an appropriate balance between the seemingly competing goals of improving speech intelligibility (strong noise reduction) and maintaining environmental sound awareness (weaker noise reduction).

Historically, this has not represented a challenge simply because existing noise reduction has not been very effective. The goal has always been complete isolation of the target speech and suppression of the background, but this is not achieved in practice. Instead, only modest attenuation of the background has been possible in commercially available noise reduction. However, deep learning based noise reduction has been shown to produce far greater isolation of target speech and suppression of the background (again see Manuscript I). This new ability to perform very highly effective noise reduction carries with it this new issue addressed currently.

This paper proposes a new T-F mask, based on the IRM, which we call the ideal compressed mask (ICM). The purpose of the ICM is to produce large speech intelligibility improvements while retaining listeners' access to environmental sound recognition (ESR). It operates by scaling down T-F units along a continuum, like the

IRM. However, whereas the standard IRM assigns attenuation values between 0 and 1, the ICM compresses the range of possible attenuation values by limiting the lower bound to a number greater than 0. Thus, no T-F units are discarded, even if they are utterly dominated by noise. But even though the ICM does not eliminate the nosiest T-F units, it still attenuates them the most. Although the current implementation was similar to the IRM, with its continuous attenuations, the ICM could also be implemented more similarly to the IQM, with discrete attenuations.

Other studies have also deliberately retained noise or added noise to speech that has been processed using T-F masking. This was done to improve target-speech intelligibility (Cao *et al.*, 2011), to balance the trade-off between intelligibility and sound quality (Brons *et al.*, 2012), or to help maintain the overall sound quality of the stimulus by limiting the introduction of "musical" noise in the signal (Anzalone *et al*., 2006). In contrast to these approaches designed to only impact the target speech, the purpose of retaining some level of background sound through T-F masking in the present study was to provide access to ESR while improving speech intelligibility.

Previous work has shown that deep neural networks (DNNs) are capable of estimating the IRM with accuracy (Wang *et al.*, 2014) sufficient to produce high speech intelligibility (e.g., Healy *et al.*, 2015, 2017, 2019; Chen *et al.*, 2016; Zhao *et al.*, 2018). This study therefore seeks to identify the most appropriate DNN training target (the best ICM version) for striking the optimal balance between speech recognition and environmental sound identification. Because the current study involves a dual task

paradigm in which the speech and the environmental sound are both signal and masker, the term “speech-to-background ratio” (SBR) will be used in place of SNR.

II. METHOD

A. Subjects

Two groups of subjects were recruited. The first group consisted of 10 paid HI listeners who were representative of typical hearing aid patients from The Ohio State University Speech-Language-Hearing Clinic. They ranged in age from 21 to 71 years (mean = 54), and five were female. All HI listeners were bilateral hearing aid users with sensorineural hearing loss. Pure-tone audiometry (ANSI, 2004, 2010) was used to confirm hearing status on day of test. Hearing-impaired listeners’ pure-tone-average thresholds (PTAs), equal to mean audiometric thresholds at 500, 1000, and 2000 Hz averaged across ears, ranged from 33 to 79 dB hearing level (HL) with a mean of 44. Hearing losses ranged from mild to profound. HI subjects were numbered in order of ascending PTA, with higher subject numbers corresponding to more mid-frequency hearing loss. Figure 4.1 displays audiograms for the 10 HI listeners, arranged from left to right and top to bottom by subject number. Each audiogram also shows the corresponding listener’s age and sex.

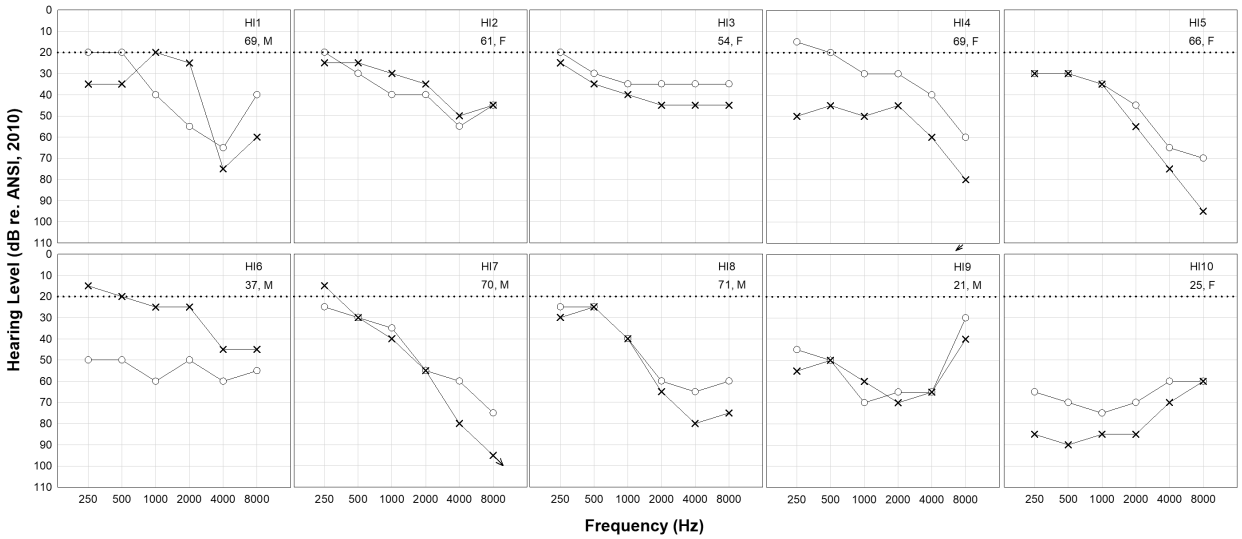

Figure 4.1. Audiometric pure-tone air-conduction thresholds for the 10 listeners with hearing impairment. Listeners are numbered in order of increasing degree of mid-frequency hearing loss. Right-ear thresholds are represented by circles, and left-ear thresholds are represented by X's. An arrow attached to a symbol indicates no response was given at the limits of the audiometer. The normal-hearing limit of 20 dB HL is marked by a horizontal dotted line in each panel. Subject numbers, ages in years, and sexes are also given.

The second group of listeners was composed of 12 NH subjects, defined as having pure-tone thresholds of 20 dB HL or better at octave frequencies from 250 to 8000 Hz on day of test (ANSI, 2004, 2010). Normal-hearing listeners' ages ranged from 19 to 20 years old (mean = 19.3). Ten were female, and two were male. Because the selection criteria for this group targeted ideal auditory processing abilities, older adults were not recruited for the NH listener group. Normal-hearing subjects received either a monetary incentive or course credit at The Ohio State University for participating. All listeners were native speakers of English, and none had any previous experience with the sentence materials employed in this experiment. Informed written consent was given to participate in this study, which was approved by The Ohio State University Institutional Review Board.

B. Stimuli

The stimuli were sentences from the Hearing in Noise Test (Nilsson *et al*., 1994) mixed with environmental sounds. The standard recording of the HINT was used, and so all sentences were produced by the same adult male talker, who was a native speaker of General American English. The same environmental sounds used in Manuscript 2 were used in the present study. These sounds were drawn from the Database of Environmental Sounds for Research Activities (Gygi and Shafiro, 2010), the Sound Effects Recognition Test (Finitzo-Hieber *et al.*, 1980), a commercial stock media website (Pond5.com), and field recordings made by the first author. On average, both HI and NH listeners achieve approximately 98% correct ESR with this set of 25 sounds when the SBR is sufficiently low (Manuscript 2). The sounds are continuous in nature and sufficiently long to overlap with an entire sentence. See Fig. 4.2 for the 25 environmental sounds used.

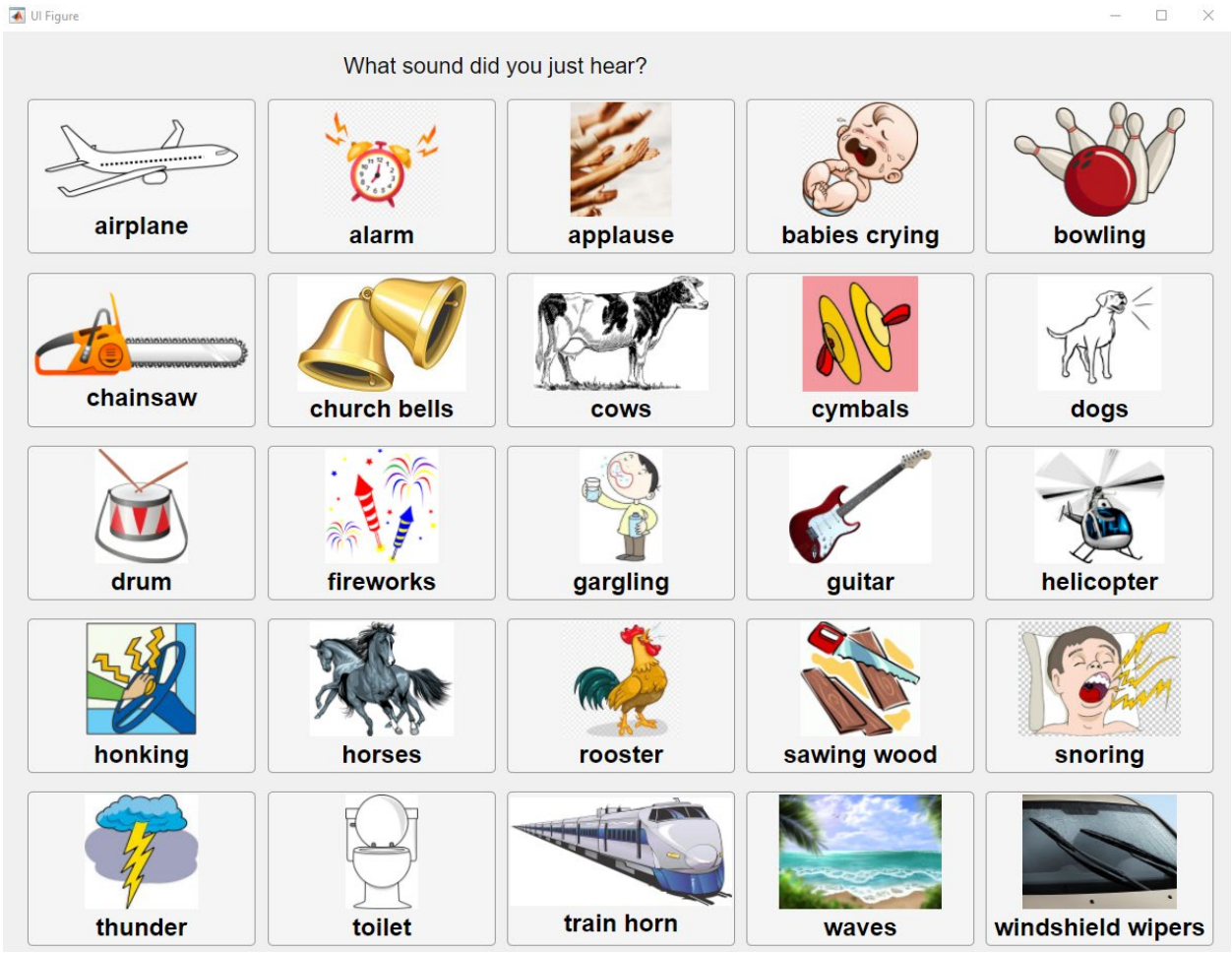

Figure 4.2. The graphical user interface for responding to environmental sounds. The 25 environmental sounds are arranged in a 5 x 5 grid in alphabetical order with pictures to facilitate listeners' visual search for the intended response.

Each sentence was mixed with a single randomly selected environmental sound at -17 dB SBR. Based on the findings of Manuscript 2, predicted speech intelligibility at this SBR was low (approximately 7% and 22% correct for HI and NH listeners, respectively), whereas predicted ESR was at ceiling (approximately 98% correct for both groups). At this SBR, therefore, speech intelligibility is severely degraded and in need of improvement while environmental sounds are readily accessible.

The onset and offset of the sentence and environmental sound in each mixture were aligned to limit asynchrony or fringe cues (Bacon and Grantham, 1992; Darwin,

1981, 1984; Oxenham and Dau, 2001; Rasch, 1978). After the RMS amplitude of the environmental sound was rescaled to be 17 dB higher than that of the sentence, the two signals were mixed together.

C. Time-frequency mask description

Let *X* and *N* represent the speech and environmental sound signals, respectively. The standard IRM is defined as:

$$IRM = \frac{S(X)}{S(X) + S(N)}$$

where *S*(.) represents the magnitude short-time Fourier transform (STFT) of a signal. This function has an output range of 0 to 1, inclusive: When no speech energy is present in a T-F unit, the IRM = 0 (i.e., -∞ dB gain, or full attenuation), and when no environmental sound energy is present, the IRM = 1 (i.e., 0 dB gain, or no attenuation). The IRM can take any value along the continuum from 0 to 1, with T-F units having a higher SBR being attenuated less and those having a lower SBR being attenuated more.

The ideal compressed mask (ICM) is defined as:

$$ICM = c \times \left[\frac{S(X)}{S(X) + S(N)}\right] + 1 - c$$

where *c* represents a compression factor between 0 and 1, inclusive. This compression factor limits the output range of the ICM to between (1 – *c*) and 1, inclusive. For example, if *c* = 1, then there is no function compression, and the ICM is simply the full IRM with an output range of 0 to 1 (i.e., -∞ to 0 dB gain). If *c* = 0.9, then the output range of the function is compressed to between 0.1 and 1, with units that are completely

dominated by the environmental sound being multiplied by 0.1 (-20 dB gain) and units that are completely speech dominated being multiplied by 1 (0 dB gain). And if $c = 0$, then the ICM function becomes fully compressed, and all units are multiplied by 1 (0 dB gain), regardless of their SBR, resulting in no change to the unprocessed mixture. Thus, the higher the value of c, the more attenuation occurs for those T-F units having low SBRs (i.e., dominated by the environmental sound).

Six compression levels (values of c) were selected for each listener group. These compression levels will be referred to by the maximum amounts of attenuation that can occur in their respective ICMs, in dB. For example, if $c = 0.99$, the output range of the ICM is between 0.01 and 1, meaning it can attenuate T-F units up to 40 dB. The compression levels selected for the HI listeners had maximum-attenuation values of 0, 20, 25, 30, 35, and ∞ dB. Again, note that a maximum-attenuation value of 0 dB simply results in the original, unprocessed mixture whereas a maximum-attenuation value of ∞ dB corresponds to the standard IRM, which fully attenuates any T-F units with no speech energy. The smaller the maximum-attenuation value, the more environmental sound energy is retained. Thus, the intermediate compression levels (maximum-attenuation values greater than 0 and less than ∞ dB) result in ICMs that retain different overall amounts of environmental sound energy in the processed signal. The selected levels were determined based on pilot testing that indicated that values in this range provided the best balance of speech intelligibility and ESR.

Figure 4.3 displays six curves plotting gain (negative gain is attenuation) as a function of local SBR (that in the particular T-F unit) for the six compression levels used

for the HI listeners in this study. The different line types (solid, dotted, dashed, etc.) represent different amounts of maximum attenuation, resulting from different compression factors for the ICM. The solid horizontal line along the top of the plot at 0 dB gain represents a fully compressed ICM, which does not alter the T-F units of a signal and whose output is equivalent to the unprocessed mixture. The steeply sloping dash-dot line represents gain as a function of SBR for the IRM, whose maximum attenuation (or negative gain) on a dB scale is unlimited. The four functions whose lower asymptotes lie between $-\infty$ and 0 dB gain represent four ICMs that limit attenuation to either 20, 25, 30, or 35 dB. As shown in Fig. 4.3, they behave similarly to the uncompressed IRM for T-F units with higher SBRs in that they cause very little attenuation under these conditions. For T-F units with lower SBRs, however, these ICMs will only attenuate down to a specified maximum amount, as indicated by their lower asymptotes.

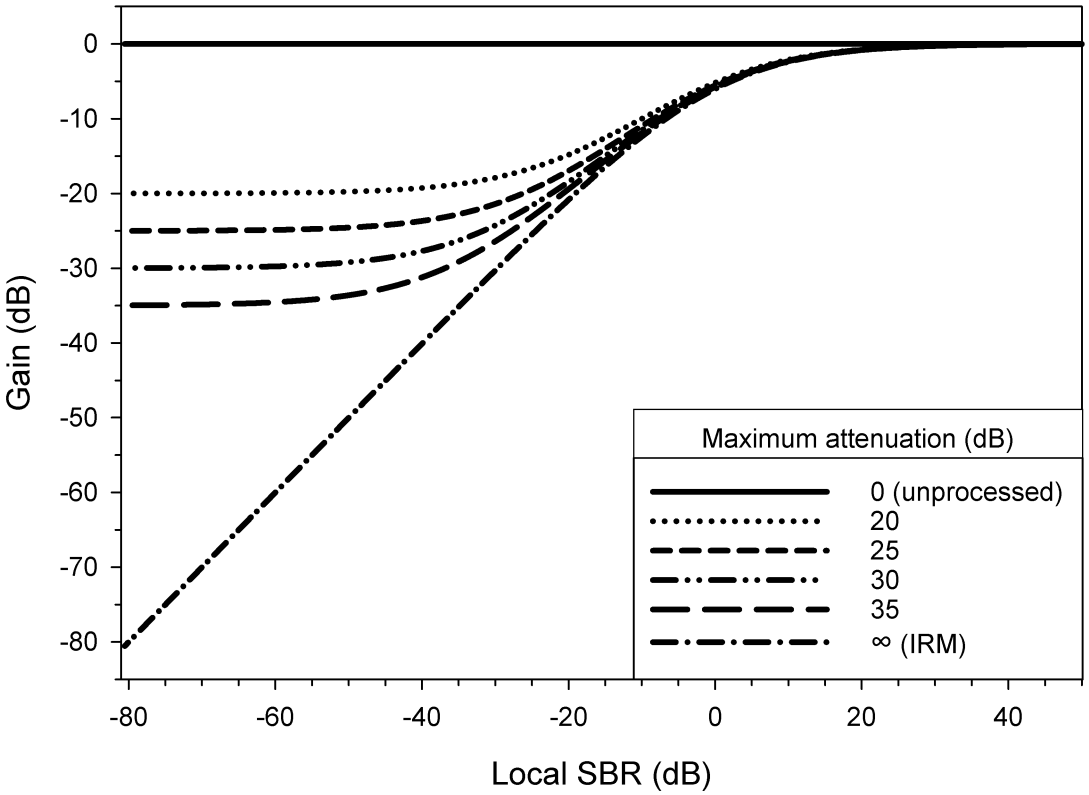

Figure 4.3. Gain as a function of local SBR for six ideal compressed masks with the six levels of compression, which were tested on HI listeners. The different line styles represent ideal compressed masks with different maximum-attenuation values.

The selected compression levels for the NH listeners had maximum-attenuation values of 0, 10, 25, 40, 55, and ∞ dB. The results of Manuscript 2 and pilot testing for this study indicated that NH listeners can perceive both speech and environmental sounds over a larger range of ICM maximum-attenuation values than HI listeners can. Accordingly, a larger range of intermediate compression levels were selected for the NH listeners in order to widen the search area for an optimal value. Note that the same unprocessed (0 dB maximum attenuation) and IRM (∞ dB maximum attenuation) control conditions are present for both the HI and NH listeners. Figure 4.4 displays six functions of gain vs. local SBR representing the six ICMs tested on the NH listeners in this study.

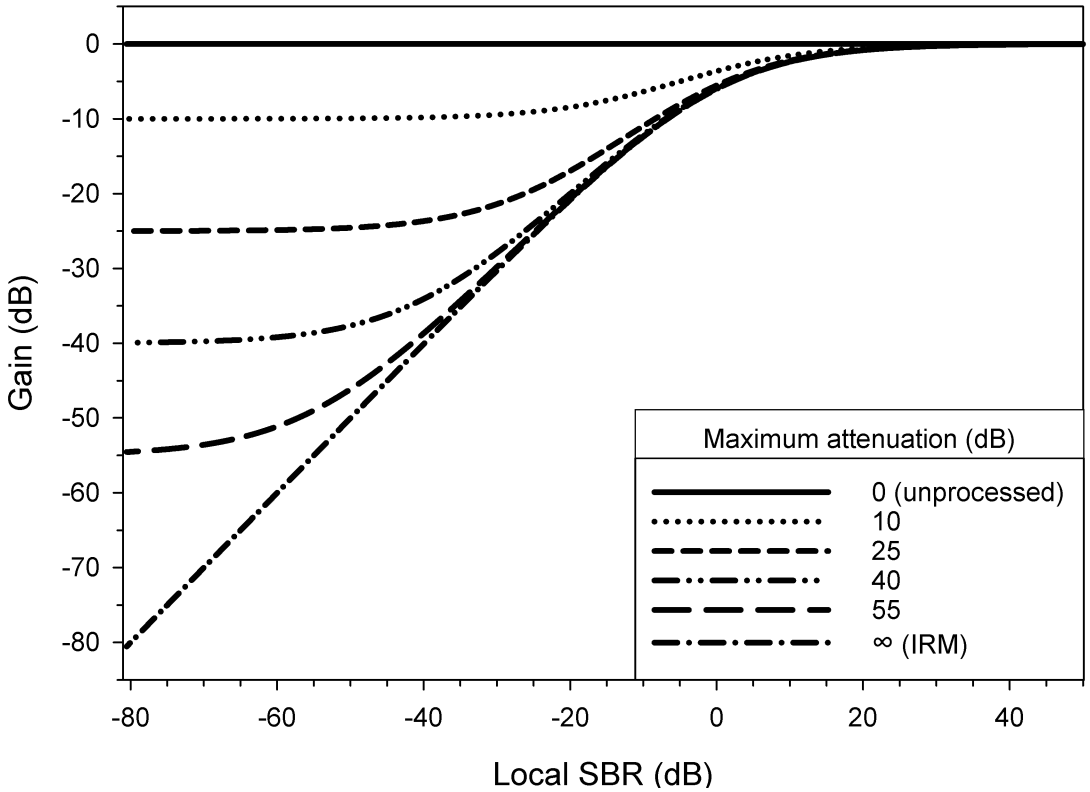

Figure 4.4. As Fig. 4.3, but plotting the ICM compression levels for the NH listeners.

To implement the various ICMs, an 882-point fast Fourier transform on Hann-windowed frames of length 20 ms with 10-ms overlap was applied to the mixture stimuli, which generated their T-F representations. Because all stimuli had an audio sampling rate of 44.1 kHz, the 882 frequency channels in the resulting T-F representations had a resolution of 25 Hz per channel. The ICM for the mixture was then point-wise multiplied by the magnitude STFT of the mixture to generate the processed magnitude STFT for each stimulus. The time-domain signal was then resynthesized through inverse STFT using the ICM-processed magnitude and the mixture-signal phase. Finally, each mixture was scaled to the same root mean square (RMS) amplitude.

D. Procedure

Each subject heard 180 stimuli blocked by six ICM compression level conditions, with 30 stimuli per condition. The conditions were presented in a random order for each subject, whereas the sentences were presented in the same fixed order for each subject. Again, the environmental sound for each experimental trial was chosen randomly with replacement. Thus, a listener may have been presented with the same sound more than once or not at all in a condition.

The experiment took place inside a double-walled audiometric booth, with the listener seated in front of a computer monitor and mouse. Listeners were instructed to attend to both the speech and the environmental sound and perform two tasks on each trial: First, repeat the sentence out loud, guessing if unsure; then, use the mouse to click on the labeled picture representing the background sound in the stimulus (Fig. 4.2), also guessing if unsure. Listeners were instructed to give their full response to the speech recognition task before clicking on their response to the ESR task to ensure that the playback of the subsequent stimulus did not begin while they were still speaking. This potential pitfall, which all listeners avoided, was a result of the self-paced nature of the experiment: each stimulus was presented automatically 300 ms after the response to the ESR task was received by the software. The experimenter, who was seated in the booth with the listener, recorded the number of words correctly reported in each sentence. For a word to be scored as correct, it had to be repeated exactly, apart from article variations (a/the) and verb tense (is/was, are/were, and has/had). The custom MATLAB application

generated a new stimulus for each trial and recorded listeners' responses to the ESR task. The test conditions were not made known to the listener or experimenter during testing.

Signals were played from a Windows PC using an RME Fireface UCX audio interface (Haimhausen, Germany), amplified using a Mackie 1202-VLZ mixer (Woodinville, WA), and presented diotically through Sennheiser HD 280 Pro headphones (Wedemark, Germany). For the NH listeners, the presentation level was set to 65 dBA in each ear. For the HI listeners, who were tested without their hearing aids, NAL-RP gains (Byrne *et al*., 1990) were added to this 65 dBA presentation level, to facilitate audibility of the stimuli. A RANE DEQ 60L digital equalizer (Mukilteo, WA) provided these gains, as described in Healy *et al*. (2015). The sound pressure level following NAL-RP amplification did not exceed 100 dBA for any participant, and all levels were verified using a sound-level meter and flat-plate coupler (Larson Davis models 824 and AEC 101, Depew, NH).

Each participant completed a three-stage familiarization before beginning formal testing. Stage 1 involved becoming familiar with the 25 isolated environmental sounds. Participants heard each of the 25 sounds one at a time, in a random order, and were instructed to click on the labeled picture corresponding to the sound after each stimulus presentation. Playback of the next sound occurred 300 ms after the click response to the current sound was received. Feedback was provided for any incorrect responses given during this stage. The set of 25 individual sounds was repeated until the listener could identify all of them without assistance from the experimenter. On average, 1.5 repetitions of this familiarization process were necessary, but up to three were permitted. One

potential HI subject failed to learn the sound labels after three repetitions, so she was dismissed from the experiment.

In stage 2 of familiarization, listeners practiced the speech recognition task and acquainted themselves with the target talker's voice. Listeners heard seven HINT sentences in quiet and were instructed to repeat back what they heard after each one. Listeners clicked a button on the screen when they were ready for the next sentence.

In familiarization stage 3, listeners practiced the dual task paradigm for the formal experiment by listening to 30 stimuli, each consisting of a HINT sentence mixed with an environmental sound at -17 dB SBR and processed using an ICM. There were five practice stimuli in each condition. For the HI subjects, the six maximum-attenuation conditions were, in order: 25, 30, 20, 35, 0 and ∞ dB. For the NH listeners, the maximum-attenuation conditions were 40, 25, 55, 10, ∞, and 0 dB, in that order. Practice conditions in this stage were arranged in an order that generally increased in terms of dual-task performance difficulty. Listeners were instructed to repeat back as much of each sentence as possible and then click on the picture representing the background sound, taking their best guess if unsure. The 37 HINT sentences used for familiarization were distinct from the 180 used for testing. Following the three familiarization stages, formal testing began, as described above. In total, each experimental session lasted approximately 45 minutes.

III. RESULTS AND DISCUSSION

Figure 4.5 displays the HI group-mean percent-correct speech intelligibility and ESR performance in each of the ICM conditions. Solid black circles represent speech-intelligibility scores, and open circles indicate ESR scores. Error bars are 95% confidence intervals. In the unprocessed condition (0 dB maximum attenuation), where speech and environmental sounds were presented at -17 dB SBR without additional processing, ESR reached 96.7% correct. This ESR score replicates the value extrapolated from the data obtained in Manuscript 2 (97.6% correct) using a different group of HI listeners and the same divided attention task. Unprocessed speech intelligibility for the HI listeners was 2.4% correct, which is also similar to the score predicted by the fitted curve for the HI group in Manuscript 2 (6.6%). These similarities between extrapolated and actual data were obtained despite the greater degree of hearing loss observed in the current HI participants (mean PTA = 44 dB HL), relative to those in Manuscript 2 (mean PTA = 30 dB HL). Speech intelligibility improved monotonically as the amount of ICM compression decreased (i.e., as maximum attenuation increased), rising to 79, 91, 95, 98 and finally 99% correct for the uncompressed ICM, which constitutes the standard IRM designed to maximize intelligibility. Even HI10, the listener with the most severe hearing loss in this study, achieved 100% correct speech intelligibility in the IRM condition despite only correctly recognizing one out of a possible 160 words (0.6% correct) in the unprocessed condition, amounting to a processing benefit of 99.4 percentage points. The HI group-mean intelligibility benefit of uncompressed (iIRM) processing was 97 percentage points.

However, this benefit to speech intelligibility came at the expense of ESR, which fell from 96.7% correct in the unprocessed condition to 8.3% correct in the IRM-processed condition, corresponding to a processing-induced ESR loss of 88.4 percentage points. Fig. 4.5 shows that ESR decreased monotonically as maximum attenuation increased (i.e., as the level of compression in the ICM decreased), falling from 96.7 to 90, 89, 77, 73, and finally 8.3% correct. This negative association between maximum-attenuation value and ESR is not surprising because the ICM is intended to most attenuate those T-F units with the highest relative environmental sound energy.

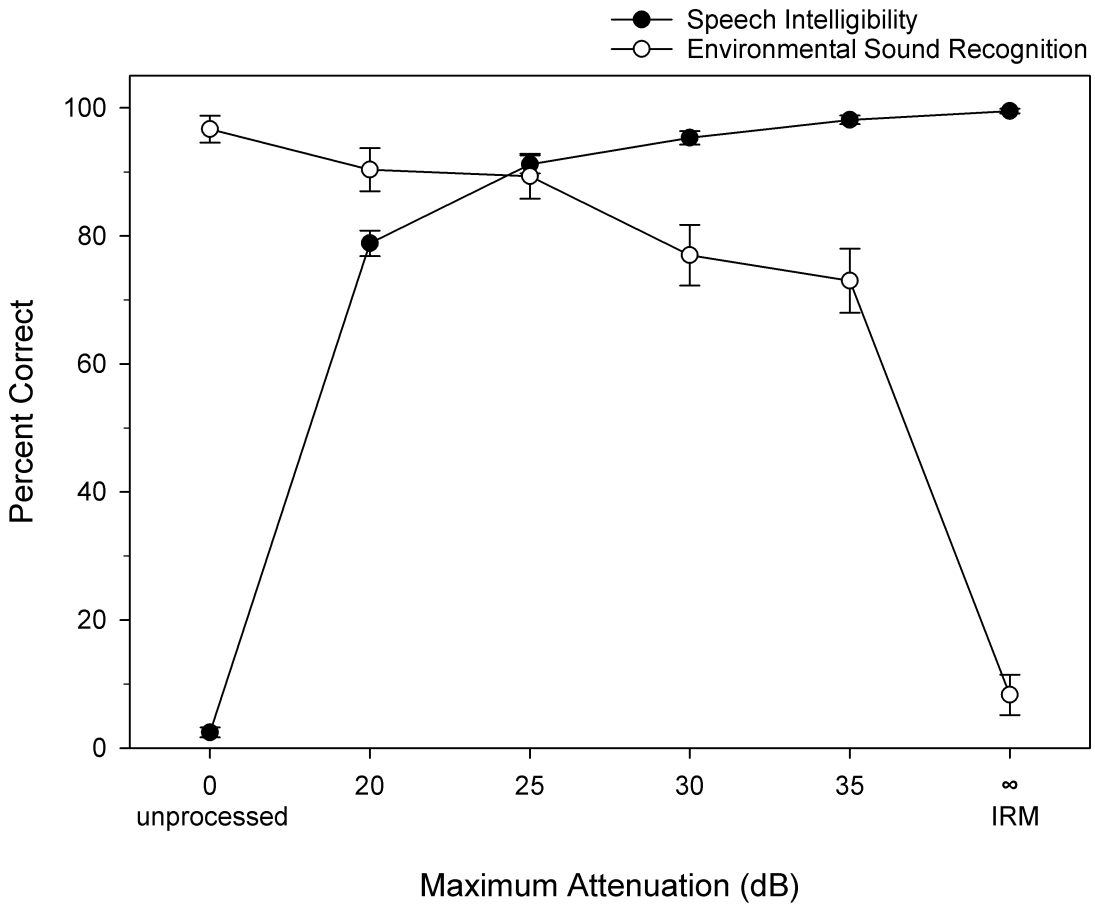

Figure 4.5. Group-mean percent-correct speech intelligibility (filled circles) and environmental sound recognition (open circles) for the 10 hearing-impaired listeners in the present study at six different maximum-attenuation levels for the ideal compressed mask. Error bars indicate 95% confidence intervals. The maximum-attenuation level of 0 dB corresponds to the unprocessed sentence-sound mixture. The maximum-attenuation level of ∞ dB corresponds to the standard (uncompressed) ideal ratio mask. The speech-to-background ratio, prior to ICM processing, was -17 dB for all stimuli.

Evident in Fig. 4.5 is that each of the four intermediate ICMs (maximum attenuation levels of 20, 25, 30, and 35 dB) provides better speech intelligibility than the unprocessed mixture (ICM maximum attenuation = 0 dB) and better ESR than the uncompressed IRM (maximum attenuation = ∞ dB). In other words, each intermediate ICM increases intelligibility from the unprocessed baseline without reducing ESR to the extent that the standard IRM does. As such, the ICM successfully improves intelligibility without completely sacrificing ESR at all four compression levels tested on HI listeners. The next question is, which of these four compression levels provides the most optimal balance of intelligibility and ESR for HI listeners? As shown in Fig. 4.5, combined intelligibility and ESR performance is highest when maximum attenuation is set to 25 dB. At this compression level, both intelligibility and ESR fall short of ceiling performance, but the sum of their combined percent-correct scores is higher in this condition than it is in any of the other five conditions tested (91% correct intelligibility + 89% correct ESR = 180 percentage points). Thus, when speech and ESR are regarded as equally important, the maximum-attenuation level of 25 dB yields the most optimal balance of speech intelligibility and ESR for these HI listeners.

However, if speech intelligibility is prioritized over ESR, as is often the case in many everyday listening situations, then setting the maximum-attenuation level to a different value may be appropriate. For example, when maximum attenuation is set to 35 dB, speech intelligibility is even higher than the observed score for the “optimal” maximum-attenuation value of 25 dB (98.1 vs 91.2% correct) while ESR remains

substantially higher than what the IRM provides (73 vs 8.3% correct). Many individuals may be willing to sacrifice 16.3 percentage points of ESR to gain 6.9 percentage points of speech intelligibility by shifting the maximum-attenuation value from 25 to 35 dB, even though this decreases the combined intelligibility + ESR score by 9.4 percentage points.

Interestingly, even though the ESR performance score of 8.3% correct in the uncompressed IRM condition is comparatively low, an exact binomial test revealed that this score is significantly higher than the 4% correct predicted by chance ($p = 0.0005$). This finding suggests that the IRM may retain enough small traces of environmental sounds to allow HI listeners to recognize them with greater-than-chance accuracy. This lack of complete removal of environmental sounds by the IRM may be explained by its finite temporal and frequency resolution as well at its use of the mixture phase (which is dominated by the environmental sound) to reconstruct the processed signal.

Figure 4.6 displays group-mean speech intelligibility and ESR scores in the six processing conditions for the NH listeners. As in Fig. 4.5, intelligibility is represented by solid circles, and ESR is represented by open circles. Note that three of the maximum-attenuation values for the NH listeners are different from the values used for the HI listeners. As with the HI listeners, all stimuli were mixed at -17 dB SBR prior to ICM processing.

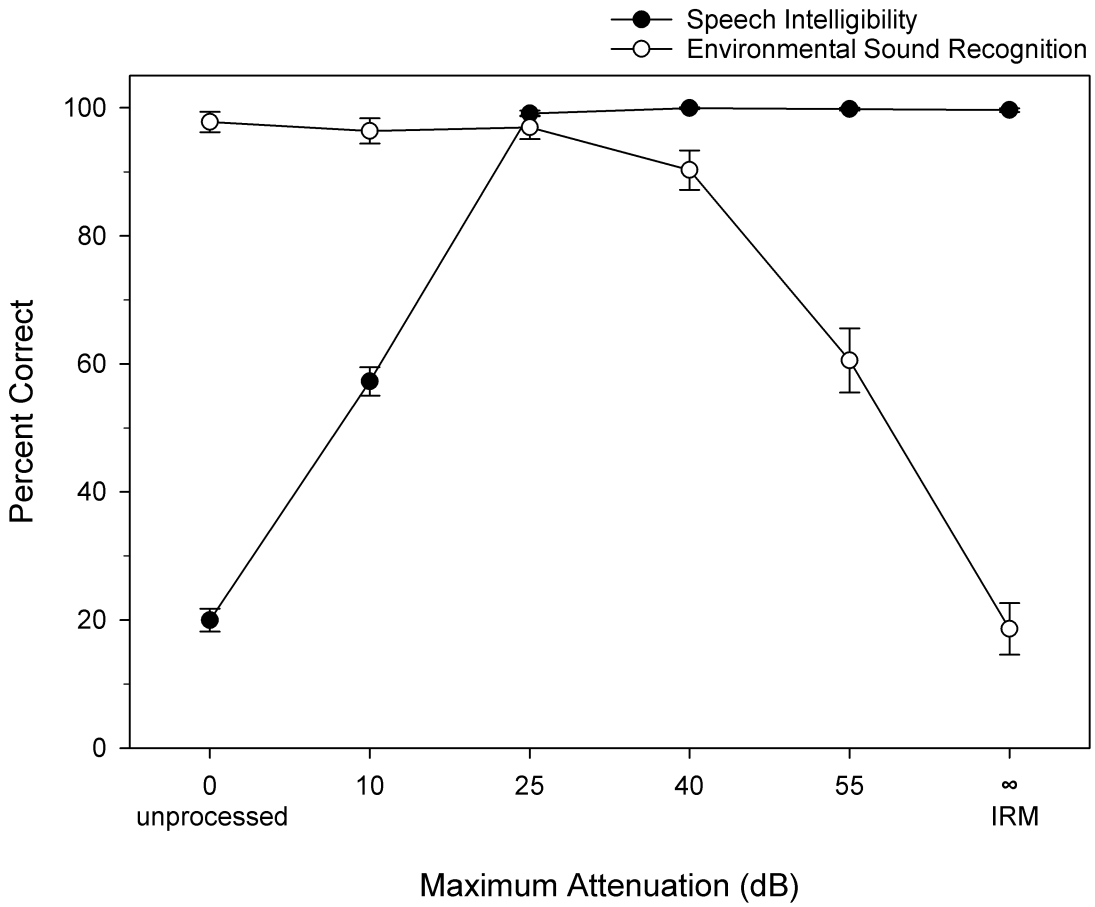

Figure 4.6. As Fig. 4.5, but for the normal-hearing listeners. Note the different set of maximum-attenuation values used for the normal-hearing listeners.

In the unprocessed condition (0 dB maximum attenuation), speech intelligibility for the NH listeners was 20.0% correct, which was within 2 percentage points of the score predicted by the results of Manuscript 2. Furthermore, ESR was 97.8% correct in this condition, which closely matched the predicted score.

In the IRM condition, the NH listeners correctly recognized 18.6% of the environmental sounds, which is significantly higher than the 4% correct predicted by chance ($p < 0.000001$). Again, this indicates that although the IRM substantially reduces ESR, it does not completely eliminate all traces of environmental sounds from the

mixture, and it appears that the NH listeners were even more sensitive to those aspects of environmental sounds that remained after IRM processing than the HI listeners were.

As observed with the HI listeners, higher maximum-attenuation values resulted in improved speech intelligibility and reduced ESR for the NH listeners, again illustrating the trade-off that exists between these two tasks. Speech intelligibility reached the performance ceiling (over 99% correct) at 25 dB maximum attenuation. ESR was also very high at this maximum-attenuation level with a score of 96.9% correct, which was within 1 percentage point of the ESR score observed in the unprocessed condition (97.8% correct). Since 25 dB maximum attenuation was the only condition tested where both speech intelligibility and ESR were at or very near their performance ceilings, this condition is considered the most optimal for balancing the trade-off between these two tasks for NH listeners.

Among the different maximum-attenuation values tested, the optimal value for achieving the highest combined intelligibility + ESR score was the same for both HI and NH listeners: 25 dB. To obtain a more exact measure of each group's optimal maximum attenuation value, a cumulative normal function was fit to each data series in Figs. 4.5 and 4.6 using the quickpsy package (Linares and López-Moliner, 2016) in R 4.1.3 (R Core Team, 2022). In order to continue using a dB scale and avoid infinite values, the IRM condition was excluded from the data used to fit the curves. In each function, the explanatory variable was maximum attenuation, and the response variable was binary data for either words or environmental sounds correctly reported. The software calculated the lapse rate (i.e., the probability of an incorrect response, independent of stimulus level)

as a free parameter. The guess rate was set to 0.04 for the ESR functions, reflecting chance performance for a closed-set task with 25 possible responses. The optimal maximum-attenuation value for achieving the highest combined intelligibility + ESR performance for each listener group was calculated as the intersection between the fitted curves for the two tasks. The crossover points for these curves were 23.4 and 23.3 dB maximum attenuation for the HI and NH groups, respectively. The optimal maximum-attenuation values for the ICM were therefore highly similar across listener types, and highly similar to that observed without curve fitting, despite the exclusion of the infinite attenuation condition.

In order to recognize speech and environmental sound simultaneously, the auditory system of the listener must be able to independently resolve the two signals in the mixture. In other words, unless the spectro-temporal resolution of the auditory system is sufficiently acute to glimpse the two signals in the mixture, the dominant signal will mask the weaker one. Because listeners with sensorineural HI often have broad auditory tuning and perhaps poor temporal resolution stemming from limited audible bandwidth and listening at low sensation levels (see Moore, 2007), they are less able to resolve the speech-sound mixture into T-F units that are small enough to contain glimpses of only a single sound source, and instead more of their available units contain a mixture of speech and environmental sound. Because of this, when concurrent speech and environmental sounds are mixed in normal fashion without further processing (as in Manuscript 2), the sounds are more likely to interfere with each other in an impaired auditory system than in a normal one, resulting in a narrower range of SBRs that allow high performance for both

speech intelligibility and ESR (as seen in Manuscript 2). However, the results of this study suggest that when resolution is dictated not by the auditory system and instead by processing in to T-F units, the same relationship between local SBR and attenuation provides the best conditions for glimpsing two signals for both HI and NH listeners. This is true despite the fact that the processed signal is then delivered to the impaired system. This result holds promise for future noise-reduction processing schemes intended for HI listeners.

In a deep learning based noise reduction system that has been trained to estimate the SBR of each T-F unit, adjusting the maximum-attenuation value for the ICM would be a simple matter. This parameter could be manipulated by the user, either through a smartphone app or a switch directly on the device, to control the amount of noise reduction performed by the algorithm. When more environmental sound awareness is desired, such as when no speech is present, the maximum-attenuation value can be decreased; when environmental sound awareness is not needed and speech is the sole signal of interest, the maximum-attenuation value may be increased until it reaches the IRM. Alternatively, the maximum-attenuation value could be adjusted automatically based on an algorithm that analyzes the listening environment and predicts the listener's desired maximum-attenuation value for the ICM. Such AI-based scene identification and settings adjustment is already implemented in commercial hearing aids (e.g., Starkey AI; Hicks, 2020). It many cases, it may be possible to adjust the maximum attenuation value in such a way that both high speech intelligibility and high ESR are provided, which represents a transformational shift in overall approach to noise reduction for HI listeners.

IV. CONCLUSIONS

1. As demonstrated in previous studies, the standard IRM delivered vastly improved speech intelligibility. However, this increased intelligibility was accompanied by a significant cost to ESR.
2. When the output range of the IRM was compressed such that its maximum attenuation was limited, it still delivered very large improvements to speech intelligibility, even rivaling the performance of the uncompressed IRM.
3. Under conditions of various levels of IRM compression, both HI and NH listeners demonstrated higher speech intelligibility than was observed in the unprocessed condition while also demonstrating higher ESR performance than was observed for the standard uncompressed IRM.
4. The optimal level of IRM compression for achieving the highest combined level of speech intelligibility and ESR was virtually the same for HI and NH listeners.
5. Future deep learning based noise reduction algorithms that have been designed to adjust the level of maximum attenuation by the estimated T-F mask may provide a better balance of speech intelligibility and ESR for listeners than an algorithm that targets clean speech and complete suppression of the background.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and other Communication Disorders (R01 DC015521 and F32 DC019314).

REFERENCES

ANSI (**2004**). S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (Acoustical Society of America, New York).

ANSI (**2010**). S3.6, *American National Standard Specification for Audiometers* (Acoustical Society of America, New York).

Anzalone, M. C., Calandruccio, L., Doherty, K. A., & Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480.

Bacon, S. P., & Grantham, D. W. (**1992**). "Fringe effects in modulation masking," J. Acoust. Soc. Am. **91**, 3451-3455.

Bell, B. The Psychological/Social Impact of Cochlear Implants [thesis]. Rochester, NY: Rochester Institute of Technology Scholar Works; **2005**;138. Available at: https://scholarworks.rit.edu/theses/8088/.

Brons, I., Houben, R., & Dreschler, W. A. (**2012**). "Perceptual effects of noise reduction by time-frequency masking of noisy speech," J. Acoust. Soc. Am. **132**, 2690-2699.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Byrne, D., Parkinson, A., and Newall, P. (**1990**). "Hearing aid gain and frequency response requirements for the severely/profoundly hearing impaired," Ear Hear. **11**, 40–49.

Cao, S., Li, L., & Wu, X. (**2011**). "Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise," J. Acoust. Soc. Am. **129**, 2227-2236.

Chen, J., Wang, Y., Yoho, S. E., Wang, D. L., and Healy, E. W. (**2016**). "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," J. Acoust. Soc. Am. **139**, 2604–2612.

Ciorba, A., Bianchini, C., Pelucchi, S., & Pastore, A. (**2012**). "The impact of hearing loss on the quality of life of elderly adults. Clinical interventions in aging," **7**, 159.

Darwin, C. J. (**1981**). "Perceptual grouping of speech components differing in fundamental frequency and onset-time. The Quarterly Journal of Experimental Psychology Section A, **33**, 185-207.

Darwin, C. J. (**1984**). "Perceiving vowels in the presence of another sound: Constraints on formant perception," J. Acoust. Soc. Am. **76**, 1636–1647.

Dillon, H. (**2012**). *Hearing Aids*, 2nd ed. (Boomerang, Turramurra, Australia), p. 232.

Finitzo-Hieber, T., Gerling, I. J., Matkin, N. D., & Cherow-Skalka, E. (**1980**). "A sound effects recognition test for the pediatric audiological evaluation," Ear Hear. **1**, 271-276.

Gygi, B., & Shafiro, V. (**2010**). "Development of the database for environmental sound research and application (DESRA): Design, functionality, and retrieval

considerations," EURASIP J. Audio Speech Mus. Process., 1-12. https://doi.org/10.5281/zenodo.2622626

Harris, MS., Boyce, L., Pisoni, D. B., Shafiro, V., & Moberly, A. C. (**2017**). "The relationship between environmental sound awareness and speech recognition skills in experienced cochlear implant users," Otology and Neurotology, **38**, e308–e314.

Healy, E. W., & Vasko, J. L. (**2018**). "An ideal quantized mask to increase intelligibility and quality of speech in noise," J. Acoust. Soc. Am. **144**, 1392-1405.

Healy, E. W., Delfarah, M., Johnson, E. M., & Wang, D. L. (**2019**). "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," J. Acoust. Soc. Am. **145**, 1378-1388.

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., & Wang, D. L. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**, 1660-1669.

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. L. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**, 3029–3038.

Hicks, M. L. (**2020**, March 5-7). *Machine Learning and Artificial Intelligence in Healthable Hearing Technology* [Conference presentation]. American Auditory Society 47th Annual Scientific & Technology Conference, Scottsdale, AZ, United States.

Hu, G., and Wang, D. L. (**2001**). "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 79–82.

Hummersone, C., Stokes, T., and Brooks, T. (**2014**). "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation*, edited by G. R. Naik and W. Wang (Springer, Berlin), pp. 349–368.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Koning, R., Madhu, N., and Wouters, J. (**2015**). "Ideal time-frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," IEEE Trans. Biomed. Eng., **62**, 331–341.

Li, N., and Loizou, P. C. (**2008a**). "Effect of spectral resolution on the intelligibility of ideal binary masked speech," J. Acoust. Soc. Am. 123, EL59–EL64.

Li, N., and Loizou, P. C. (**2008b**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. 123, 1673–1682.

Licklider, J. C. R. & Miller, G. A. (**1951**) "The perception of speech," In: Stevens SS (ed) *Handbook of Experimental Psychology*. New York: John Wiley, pp. 1040–1074.

Linares, D., & López-Moliner, J. (**2016**). “quickpsy: An R package to fit psychometric functions for multiple groups,” The R J. **8**, 122-131.

Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL), Chaps. 5–8.

Madhu, N., Spriet, A., Jansen, S., Koning, R., and Wouters, J. (**2013**). “The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: Application to auditory prostheses,” IEEE Trans. Audio Speech, Lang. Process, **21**, 63–72.

Mick, P., Foley, D., Lin, F., Pichora-Fuller, M. K. (**2018**). “Hearing difficulty is associated with injuries requiring medical care," Ear Hear. **39**, 631–644.

Moore, B. C. J. (**2007**). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*, 2nd ed. (Wiley, Chichester, UK), pp. 45–91.

Narayanan, A., and Wang, D. L. (**2013**). “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7092–7096.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (**1994**). “Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise,” J. Acoust. Soc. Am. **95**, 1085-1099.

Oxenham, A. J., & Dau, T. (**2001**). Modulation detection interference: Effects of concurrent and sequential streaming,” J. Acoust. Soc. Am. **110**, 402–408.

R Core Team. (**2022**). R: A language and environment for statistical computing. R Foundation for Statistical Computing," http://www.R-project.org/

Rasch, R. A. (**1978**). "Perception of simultaneous notes such as in polyphonic music," Acta Acustica United with Acustica, **40**, 21–33.

Sinex, D. G. (**2013**). "Recognition of speech in noise after application of time-frequency masks: Dependence on frequency and threshold parameters," J. Acoust. Soc. Am. **133**, 2390–2396.

Srinivasan, S., Roman, N., and Wang, D. L. (**2006**). "Binary and ratio time-frequency masks for robust speech recognition," Speech Communication, **48**, 1486–1501.

U.S. Fire Administration, Federal Emergency Management Agency. Fire Risks for the Deaf or Hard of Hearing. December **1999**.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell MA), pp. 181–197.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., & Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336-2347.

Wang, Y. and Wang, D. L. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio. Speech Lang. Process. 21, 1381-1390.

Wang, Y., Narayanan, A., and Wang, D. L. (**2014**). "On training targets for supervised speech separation," IEEE/ACM Trans. Audio Speech Lang. Process. **22**, 1849–1858.

Zhao, Y., Wang, D. L., Johnson, E. M., and Healy, E. W. (**2018**). "A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions," J. Acoust. Soc. Am. **144**, 1627–1637.

Chapter 5. General Summary and Discussion

The three manuscripts contained here examined and discussed important considerations regarding the efficacy and viability of deep learning based noise reduction, which represents perhaps our most promising solution to the speech-in-noise problem. The first manuscript is an extension of an ongoing line of work aimed toward advancing deep-learning-based noise reduction. It involved human-subjects testing of a novel, time-domain-based algorithm that fulfills four of the most fundamental requirements for real-world implementation: talker independence, corpus independence, noise independence, and causality. The second manuscript considered a new dimension of efficacy for deep learning based noise reduction: the preservation of environmental sound recognition (ESR). Finally, the third manuscript examined a novel algorithmic scheme for achieving an optimal balance between improving intelligibility and maintaining ESR.

Our ability to increase intelligibility through deep learning based noise reduction, especially for hearing impaired (HI) listeners, has increased substantially since 2013, when the first demonstration of improved intelligibility for HI listeners was provided. Manuscript 1 tested the effect of a novel algorithm on human listeners' speech intelligibility and compared its efficacy to the seminal algorithm first described nine years ago. The stimuli, procedures, and human subject populations were essentially identical across studies, thus isolating the variables of interest, which were the algorithm used and the demands placed on it. The algorithm in the initial study was trained and

tested in highly similar conditions. Further, it featured a non-causal architecture, which prevented it from running in real time. In contrast, the new algorithm, which featured a modern attentive recurrent network design, is much more viable: It had mismatched training versus test conditions in terms of talkers, speech corpora, and noise types, and it was fully causal, thus satisfying the fundamental requirement for real-time operation. Significant algorithm benefit was observed in every condition and averaged 51 percentage points across conditions for HI listeners. Strikingly, the amount of benefit provided by the new algorithm was similar to the benefit obtained in the initial demonstration in 2013, despite the considerably more demanding conditions in which the new algorithm was tested. Thus, the new algorithm was just as efficacious as the original, but since it was causal as well as talker-, corpus-, and noise-independent, it was immeasurably more viable. Continual technological advances in deep learning based noise reduction over the past nine years have enabled this newest algorithm to provide substantial intelligibility benefit despite the systematic removal of various constraints to address viability considerations.

The second manuscript tests the limits of the human auditory system and discusses implications for future noise-reduction technology. Speech recognition and environmental sound awareness are both important aspects of the human auditory experience. However, when speech and environmental sounds occur concurrently, one can mask the other, leading to poor speech intelligibility and/or environmental sound identification. In this vein, the second manuscript demonstrates that there exists a range of speech-to-background ratios (SBRs) over which the human auditory system can

segregate and recognize speech and environmental sounds independently or simultaneously, even without the aid of binaural cues. This was accomplished by determining NH and HI listeners' ESR in the presence of concurrent speech as well as their speech reception in the presence of concurrent environmental sounds. By comparing these two performances to each other, the range of SBRs over which *both* speech recognition and ESR scores are high was calculated. It was found that both NH and HI listeners were capable of reliably recognizing both speech and concurrent environmental sounds when the SBR was optimal. The optimal SBR for the NH group (12.2 dB) was 7.1 dB higher than optimal value for the HI listeners (5.1 dB). This difference in optimal SBR values was primarily caused by the HI listeners' poorer performance on the ESR task.

Speech recognition and ESR are fundamentally different tasks that leverage different aspects of auditory perception, and it appears that speech caused a greater amount of interference on ESR for the HI listeners than for the NH listeners. This may also be due to differences in the ways NH and HI listeners use divided attention. One of the goals of Experiment 2 in the second manuscript was to establish the basic performance differences between divided and selective attention in the current task, without additional possible factors associated with aging and/or hearing loss, which is why only NH listeners participated. Future studies, however, could examine the differential effects of attention type on NH vs. HI listeners for these tasks. Only modest differences were observed across these tasks.

Deep-learning-based noise reduction can significantly improve the intelligibility of speech at low SNRs, where environmental sounds (i.e., background noise) would normally render speech unintelligible via masking. The goal of traditional noise-reduction algorithms is to output clean, noise-free speech, but this has the adverse secondary effect of reducing environmental sound awareness. The data from Manuscript 2 demonstrate that there exist SBRs where both NH and HI listeners can recognize speech and identify environmental sounds. In any situation where HI listeners desire both speech intelligibility and environmental sound awareness, the optimal SBR of 5.1 dB should be the target of noise-reduction algorithms, not the SBR corresponding to perfectly clean speech, which is ∞ dB. Of course, choosing any target SBR assumes that the deep learning algorithm will accurately deliver the signal at this target SBR. In many cases, noise remains in the algorithm-processed signal even when clean speech is the target. However, as deep learning algorithms become more advanced, they will continue to reduce estimation errors and produce outputs that better reflect their intended target.

Another aspect of balancing speech intelligibility with ESR that is worth considering is the contribution of the direct sound path into the listener's ear. This does not apply to (non-hybrid) cochlear implants since all auditory stimulation to the implanted ear comes from the device. However, with regard to hearing aids, the more open the fitting is, the greater the amount of acoustic leakage into the ear canal will occur. Thus, even if the hearing aid delivers perfectly noise-free speech, the target signal may be masked by noise that bypasses the hearing aid's noise reduction and enters the ear canal directly from the sound source. Thus, the true optimal SBR refers to the desired

SBR at the listener's ear drum resulting from the combination of hearing aid output and natural sound that enters directly into the ear canal from the outside environment. One hearing aid manufacturer has conceived a clever way to address this problem: a mechanically gated venting system. When maximum noise reduction is desired, the vent in the hearing aid coupling closes to limit the amount outside sound that enters the ear canal directly; and when minimal occlusion is desired, the vent opens to allow more sound to bypass the hearing aid and enter the ear canal.

The goal of the third manuscript was to demonstrate a viable and efficacious processing scheme capable of improving the intelligibility of noisy speech without unduly compromising environmental sound awareness. This method implemented ideal time-frequency masking for purposes of demonstration (ideal as opposed to algorithm estimated). Realistically, a hearing aid cannot have perfect knowledge of the speech vs. noise composition of an input signal that is required to generate an ideal time-frequency mask, but deep neural networks are capable of estimating time-frequency masks with a high degree of accuracy , allowing them to be acoustically similar to ideal masks (e.g., Wang *et al*., 2014) and similar in terms of their ability to produce high human intelligibility (e.g., Healy *et al*., 2019). Thus, the ideal processing used in this manuscript will serve as a proxy for advanced deep-learning-based algorithms that are trained to target the ideal processing schemes presented here.

It was found that this ICM processing was indeed capable of producing high speech intelligibility and simultaneous high recognition accuracy of concurrent environmental background sounds. Interestingly, the optimal values for such processing

were found to be highly similar across HI and NH listeners when ICM processing was employed. This result differs somewhat from that observed in Manuscript 2, where unprocessed speech and environmental sounds were employed and ideal SNRs (SBRs) were obtained. This difference may be related to the spectral and temporal resolution of the impaired vs. normal auditory systems – although other differences exist, the auditory system was tasked with dividing the sound mixture into T-F units in Manuscript 2, whereas this processing was performed for the listener through signal processing in Manuscript 3.

Taken together, these three manuscripts indicate a need to consider many separate factors when examining the efficacy and viability of deep learning based noise reduction. Efficacy and viability largely depend on the neural network that is trained to perform the speech separation task. However, the nature of that task also influences the efficacy and viability of the system since some targets may result in better outcomes for environmental sound awareness than others.

In conclusion, the data presented here have both theoretical and practical implications. From a theoretical perspective, these data provide additional insight into the ability of the normal and impaired auditory system to perceive speech and environmental sounds simultaneously. Practical implications include suggestions for modifying deep learning algorithms to target lower SBRs to preserve ESR while still delivering large benefit to speech intelligibility. Furthermore, the third manuscript presents a novel T-F masking scheme for achieving this aim.

REFERENCES

Healy, E. W., Vasko, J. L., & Wang, D. L. (**2019**). "The optimal threshold for removing noise from speech is similar across normal and impaired hearing—a time-frequency masking study,". J. Acoust. Soc. Am. ***145.*** EL581-EL586.

Cumulative References

Aniansson, G. (**1978**). "Speech intelligibility in and speech interference levels of traffic noise in hearing-impaired and normal listeners," Acta Oto-Laryngologica, **86**, 109-112.

ANSI (**2004**). S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (American National Standards Institute, New York).

ANSI (**2010**). S3.6, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).

Anzalone, M. C., Calandruccio, L., Doherty, K. A., & Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480.

Apoux, F., Carter, B. L., & Healy, E. W. (**2018**). "Effect of dual-carrier processing on the intelligibility of concurrent vocoded sentences," J. Speech Lang. Hear. Res. **61**, 2804-2813.

Ba, J. L., Kiros, J. R., and Hinton, G. E. **(2016).** "Layer normalization," arXiv:1607.06450.

Bacon, S. P., & Grantham, D. W. (**1992**). "Fringe effects in modulation masking," J. Acoust. Soc. Am. **91**, 3451-3455.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (**2014**). "Fitting linear mixed-effects models using lme4," arXiv preprint arXiv:1406.5823.

Bell, B. The Psychological/Social Impact of Cochlear Implants [thesis]. Rochester, NY: Rochester Institute of Technology Scholar Works; **2005**;138. Available at: https://scholarworks.rit.edu/theses/8088/.

Bell, B. The Psychological/Social Impact of Cochlear Implants [thesis]. Rochester, NY: Rochester Institute of Technology Scholar Works; **2005**;138. Available at: https://scholarworks.rit.edu/theses/8088/.

Blackwell, D. L., Lucas, J. W., & Clarke, T. C. (**2014**). "Summary health statistics for U.S. adults: National Health Interview Survey," 2012. National Center for Health Statistics. Vital Health Stat, **10**. Retrieved from https://www.cdc.gov/nchs/data/series/sr_10/sr10_260.pdf

Bregman, A. S. (**1990**) *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Brons, I., Houben, R., & Dreschler, W. A. (**2012**). "Perceptual effects of noise reduction by time-frequency masking of noisy speech," J. Acoust. Soc. Am. **132**, 2690-2699.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101-1109.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Busch, A. C., & Eldredge, D. (**1967**). "The effect of differing noise spectra on the consistency of identification of consonants," Language and Speech, **10**, 194-202.

Byrne, D., Parkinson, A., and Newall, P. (**1990**). "Hearing aid gain and frequency response requirements for the severely/profoundly hearing impaired," Ear Hear. **11**, 40–49.

Byrne, D., Parkinson, A., and Newall, P. (**1990**). "Hearing aid gain and frequency response requirements for the severely/profoundly hearing impaired," Ear Hear. **11**, 40–49.

Cao, S., Li, L., & Wu, X. (**2011**). "Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise," J. Acoust. Soc. Am. **129**, 2227-2236.

Chen, J., Wang, Y., and Wang, D. L. **(2015)**. "Noise perturbation improves supervised speech separation," in *Proceedings of LVA/ICA*, pp. 83–90.

Chen, J., Wang, Y., Yoho, S. E., Wang, D. L., & Healy, E. W. (**2016**). "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," J. Acoust. Soc. Am. **139**, 2604-2612.

Ciorba, A., Bianchini, C., Pelucchi, S., & Pastore, A. (**2012**). "The impact of hearing loss on the quality of life of elderly adults. Clinical interventions in aging," **7**, 159.

Cooper Jr, J. C., & Cutts, B. P. (**1971**). "Speech discrimination in noise," Journal of Speech and Hearing Research, **14**, 332-337.

Darwin, C. J. (**1981**). "Perceptual grouping of speech components differing in fundamental frequency and onset-time. The Quarterly Journal of Experimental Psychology Section A, **33**, 185-207.

Darwin, C. J. (**1984**). "Perceiving vowels in the presence of another sound: Constraints on formant perception," J. Acoust. Soc. Am. **76**, 1636–1647.

Dillon, H. (**2012**). *Hearing Aids*, 2nd ed. (Boomerang, Turramurra, Australia), p. 232.

Finitzo-Hieber, T., Gerling, I. J., Matkin, N. D., & Cherow-Skalka, E. (**1980**). "A sound effects recognition test for the pediatric audiological evaluation," Ear Hear. **1**, 271-276. https://doi.org/10.1097/00003446-198009000-00007

Finitzo-Hieber, T., Gerling, I. J., Matkin, N. D., & Cherow-Skalka, E. (**1980**). "A sound effects recognition test for the pediatric audiological evaluation," Ear Hear. **1**, 271-276.

Fletcher, H. (**1953**). *Speech and Hearing in Communication*. New York: Van Nostrand (reprinted by the Acoustical Society of America, 1995).

Fogerty, D., Carter, B. L., & Healy, E. W. (**2018**). "Glimpsing speech in temporally and spectro-temporally modulated noise," J. Acoust. Soc. Am. **143**, 3047-3057.

Goehring, T., Bolner, F., Monaghan, J. J. M., van Dijk, B., Zarowski, A., and Bleeck, S. (**2017**). "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," Hear. Res. **344**, 183–194.

Goehring, T., Chapman, J. L., Bleeck, S., & Monaghan, J. J. (**2018**). "Tolerable delay for speech production and perception: Effects of hearing ability and experience with hearing aids," Int. J of Audio., **57**, 61-68.

Goehring, T., Keshavarzi, M., Carlyon, R. P., and Moore, B. C. J. (**2019**). "Using recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants," J. Acoust. Soc. Am. **146**, 705–718.

Gygi, B., & Shafiro, V. (**2010**). "Development of the database for environmental sound research and application (DESRA): Design, functionality, and retrieval considerations," EURASIP J. Audio Speech Mus. Process., 1-12. https://doi.org/10.5281/zenodo.2622626

Harris, M. S., Boyce, L., Pisoni, D. B., Shafiro, V., & Moberly, A. C. (**2017**). "The relationship between environmental sound awareness and speech recognition skills in experienced cochlear implant users," Otology and Neurotology, **38**, e308–e314.

Harris, MS., Boyce, L., Pisoni, D. B., Shafiro, V., & Moberly, A. C. (**2017**). "The relationship between environmental sound awareness and speech recognition skills in experienced cochlear implant users," Otology and Neurotology, **38**, e308–e314.

Hawkins Jr, J. E., & Stevens, S. S. (**1950**). "The masking of pure tones and of speech by white noise," J. Acoust. Soc. Am. **22**, 6-13.

He, K., Zhang, X., Ren, S., and Sun, J. **(2016)**. "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Healy E. W., Yoho S. E., Chen J., Wang Y., and Wang D. L. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**, 1660-1669.

Healy, E. W., & Vasko, J. L. (**2018**). "An ideal quantized mask to increase intelligibility and quality of speech in noise," J. Acoust. Soc. Am. **144**, 1392-1405.

Healy, E. W., Delfarah, M., Johnson, E. M., & Wang, D. L. (**2019**). "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," J. Acoust. Soc. Am. **145**, 1378-1388.

Healy, E. W., Delfarah, M., Vasko, J. L., Carter, B. L., & Wang, D. L. (**2017**). "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," J. Acoust. Soc. Am. **141**, 4230-4239.

Healy, E. W., Johnson, E. M., Delfarah, M., and Wang, D. L. (**2020**). "A talker-independent deep learning algorithm to increase intelligibility for hearing-impaired listeners in reverberant competing talker conditions," J. Acoust. Soc. Am. **147**, 4106-4118.

Healy, E. W., Taherian, H., Johnson, E. M., and Wang, D. L. **(2021c)**. "A causal and talker-independent speaker-separation/dereverberation deep learning algorithm: Cost associated with conversion to real-time capable operation," J. Acoust. Soc. Am., **150**, 3976-3986.

Healy, E. W., Tan, K., Johnson, E. M., and Wang, D. L. **(2021b)**. "An effectively causal deep learning algorithm to increase intelligibility in untrained noises for hearing-impaired listeners," J. Acoust. Soc. Am., **149**, 3943-3953.

Healy, E. W., Vasko, J. L., & Wang, D. L. (**2019**). "The optimal threshold for removing noise from speech is similar across normal and impaired hearing—a time-frequency masking study,". J. Acoust. Soc. Am. ***145.*** EL581-EL586.

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., & Wang, D. L. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**, 1660-1669.

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. L. **(2013)**. "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am.,. **134**, 3029-3038.

Healy, E.W., Johnson, E.M., Delfarah, M., Sevich, V.A., Krishnagiri, D.S. and Wang, D. L. **(2021a)**. "Deep learning based speaker separation and dereverberation can generalize across different languages to improve intelligibility," J. Acoust. Soc. Am., **150**, 2526-2538.

Hendrycks, D., and Gimpel, K. **(2016)**. "Gaussian error linear units (gelus)," arXiv preprint arXiv:1606.08415.

Hicks, M. L. (**2020**, March 5-7). *Machine Learning and Artificial Intelligence in Healthable Hearing Technology* [Conference presentation]. American Auditory Society 47th Annual Scientific & Technology Conference, Scottsdale, AZ, United States.

Hinton, G. E., Osindero, S., and Teh, Y.-W. **(2006)**. "A fast learning algorithm for deep belief nets," Neural Comput. **18**, 1527–1554.

Hochreiter, S., and Schmidhuber, J. **(1997)**. "Long short-term memory," Neural Comput. **9**, 1735–1780.

Howard-Jones, P. A., & Rosen, S. (**1993**). "Uncomodulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. 93(5), 2915-2922.

Hu, G., and Wang, D. L. (**2001**). "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 79–82.

Humes, L. E., Lee, J. H., & Coughlin, M. P. (**2006**). "Auditory measures of selective and divided attention in young and older adults using single-talker competition," J. Acoust. Soc. Am. **120**, 2926-2937.

Hummersone, C., Stokes, T., and Brooks, T. (**2014**). "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation*, edited by G. R. Naik and W. Wang (Springer, Berlin), pp. 349–368.

Jenkins, J. J. (**1985**). "Acoustic information for objects, places, and events," In W. H. Warren, & R. E. Shaw (Eds.), Persistence and change, (pp. 115–138). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Jensen, J., and Taal, C. H. (**2016**). "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE/ACM Trans. Audio Speech Lang. Proc. **24**, 2009–2022.

Kearns, J. **(2014)**. "LibriVox: Free public domain audiobooks," Reference Reviews.

Keshavarzi, M., Goehring, T., Turner, R. E., and Moore, B. C. J. **(2019)**. "Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: A deep recurrent neural network and spectral subtraction," J. Acoust. Soc. Am. **145**, 1493–1503.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.

Kingma, D. P., and Ba, J. **(2014)**. "Adam: A method for stochastic optimization,: arXiv:1412.6980.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Klumpp, R. G., & Webster, J. C. (**1963**). "Physical Measurements of Equally Speech-Interfering Navy Noises," J. Acoust. Soc. Am. **35**, 1328-1338.

Koning, R., Madhu, N., and Wouters, J. (**2015**). "Ideal time-frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," IEEE Trans. Biomed. Eng., **62**, 331–341.

Kryter, K. D., & Williams, C. E. (**1966**). "Masking of speech by aircraft noise," J. Acoust. Soc. Am. **39**, 138-150.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (**2017**). "lmerTest package: tests in linear mixed effects models," Journal of statistical software, **82**, 1-26.

Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J.R. **(2018)**. "SDR – half-baked or well done?" arXiv:1811.02508v1.

Lewis, J. W., Wightman, F. L., Brefczynski, J. A., Phinney, R. E., Binder, J. R., & DeYoe, E. A. (**2004**). "Human brain regions involved in recognizing environmental sounds," Cerebral cortex, **14**, 1008-1021.

Li, N., and Loizou, P. C. (**2008a**). "Effect of spectral resolution on the intelligibility of ideal binary masked speech," J. Acoust. Soc. Am. 123, EL59–EL64.

Li, N., and Loizou, P. C. (**2008b**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. 123, 1673–1682.

Licklider, J. C. R. & Miller, G. A. (**1951**) "The perception of speech," In: Stevens SS (ed) Handbook of Experimental Psychology. New York: John Wiley, pp. 1040–1074.

Licklider, J. C. R., & Guttman, N. (**1957**). "Masking of speech by line-spectrum interference," J. Acoust. Soc. Am. **29**, 287-296.

Linares, D., & López-Moliner, J. (**2016**). "quickpsy: An R package to fit psychometric functions for multiple groups," The R J. **8**, 122-131.

Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL), Chaps. 5–8.

Madhu, N., Spriet, A., Jansen, S., Koning, R., and Wouters, J. (**2013**). "The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: Application to auditory prostheses," IEEE Trans. Audio Speech, Lang. Process, **21**, 63–72.

Merity, S. **(2019)**. "Single headed attention RNN: Stop thinking with your head," arXiv:1911.11423.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., ... and Wu, H. **(2017)**. "Mixed precision training," arXiv:1710.03740.

Mick, P., Foley, D., Lin, F., Pichora-Fuller, M. K. (**2018**). “Hearing difficulty is associated with injuries requiring medical care,” Ear Hear. **39**, 631–644.

Mick, P., Foley, D., Lin, F., Pichora-Fuller, M. K. (**2018**). “Hearing difficulty is associated with injuries requiring medical care," Ear Hear. **39**, 631–644.

Miller, G. A. (**1947**). “The masking of speech,” Psych. Bull. **44**, 105-129.

Monaghan, J. J. M., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C. M., and Bleeck, S. (**2017**). “Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners,” J. Acoust. Soc. Am. **141**, 1985–1998.

Moore, B. C. J. (**2007**). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*, 2nd ed. (Wiley, Chichester, UK), pp. 45–91.

Narayanan, A., and Wang, D. L. (**2013**). “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7092–7096.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (**1994**). “Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise,” J. Acoust. Soc. Am. **95**, 1085-1099.

Oxenham, A. J., & Dau, T. (**2001**). Modulation detection interference: Effects of concurrent and sequential streaming,” J. Acoust. Soc. Am. **110**, 402–408.

Panayotov, V., Chen, G., D., and Khudanpur, S. **(2015)**. "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210.

Pandey, A., & Wang, D. L. (**2021**). "Dense CNN with self-attention for time-domain speech enhancement," IEEE/ACM Trans. Audio. Speech Lang. Process. **29**, 1270-1279.

Pandey, A., and Wang, D. L. **(2020a)**. "On cross-corpus generalization of deep learning based speech enhancement," IEEE/ACM Trans. Audio. Speech Lang. Process. **28**, 2489–2499.

Pandey, A., and Wang, D. L. **(2020b)**. "Learning complex spectral mapping for speech enhancement with improved cross-corpus generalization," In INTERSPEECH, pp. 4511–4515.

Pandey, A., and Wang, D. L. **(2022)**. "Self-attending RNN for speech enhancement to improve cross-corpus generalization," IEEE/ACM Trans. Audio. Speech Lang. Process. **30**, 1374–1385.

Parasuraman R. (**1998**). "The attentive brain: Issues and prospects," In: Parasuraman R, editor. *The Attentive Brain.* Cambridge, Massachusetts: MIT Press, pp. 3–15.

Paul, D. B., and Baker, J. **(1992)**. "The design for the Wall Street Journal-based CSR corpus," In Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992.

R Core Team. (**2022**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing. http://www.R-project.org/

R Core Team. (**2022**). R: A language and environment for statistical computing. R Foundation for Statistical Computing," http://www.R-project.org/

Ramsdell, D. A. (**1978**). "The psychology of the hard-of-hearing and the deafened adult,". In H. David, & S. R. Silverman (Eds.), *Hearing and deafness*. New York: Holt, Rinehart & Winston.

Rasch, R. A. (**1978**). "Perception of simultaneous notes such as in polyphonic music," Acta Acustica United with Acustica, **40**, 21–33.

Reed, C. M., & Delhorne, L. A. (**2005**). "Reception of environmental sounds through cochlear implants," Ear Hear. **26**, 48-61.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (**2001**). "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 749–752.

Ross, M., Huntington, D. A., Hayes, A. N., & Dixon, R. F. (**1965**). "Speech discrimination of hearing-impaired individuals in noise," J. of Aud. Res. **5**, 47–72.

Santurette, S., Ng, E.H.N., Jensen, J.J., and Loong, B.M.K. (**2020**). "Oticon More clinical evidence," Oticon Whitepaper.

Sinex, D. G. (**2013**). "Recognition of speech in noise after application of time-frequency masks: Dependence on frequency and threshold parameters," J. Acoust. Soc. Am. **133**, 2390–2396.

Srinivasan, S., Roman, N., and Wang, D. L. (**2006**). "Binary and ratio time-frequency masks for robust speech recognition," Speech Communication, **48**, 1486–1501.

Stevens, S. S., Miller, J., & Truscott, I. (**1946**). "The masking of speech by sine waves, square waves, and regular and modulated pulses," J. Acoust. Soc. Am. **18**, 418-424.

Stone, M. A., & Moore, B. C. (**1999**). "Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses," Ear Hear. **20**, 182-192.

Stone, M. A., & Moore, B. C. (**2005**). "Tolerable hearing-aid delays: IV. Effects on subjective disturbance during speech production by hearing-impaired subjects," Ear Hear. **26**, 225-235.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech, Lang., Hear. Res. **28**, 455–462.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Trans. Audio. Speech. Lang. Process. **19**, 2125–2136.

Tye-Murray, N., Spehar, B., Sommers, M., and Barcroft, J. (**2016**). "Auditory training with frequent communication partners," J. Speech, Lang., Hear., Res. **59**, 871–875.

U.S. Fire Administration, Federal Emergency Management Agency. Fire Risks for the Deaf or Hard of Hearing. December **1999**.

Varga, A., and Steeneken, H. J. **(1993)**. "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Comm., **12**, 247-251.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. **(2017)**. "Attention is all you need," Advances in neural information processing systems, 30.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell MA), pp. 181–197.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., & Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336-2347.

Wang, Y. and Wang, D. L. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio. Speech Lang. Process. 21, 1381-1390.

Wang, Y., Han, K., and Wang, D. L. **(2013)**. "Exploring monaural features for classification-based speech segregation," IEEE Trans. Audio. Speech Lang. Process. **21**, 270–279.

Wang, Y., Narayanan, A., & Wang, D. L. (**2014**). "On training targets for supervised speech separation," IEEE/ACM Trans. Audio. Speech Lang. Process. **22**, 1849-1858.

Zhao, Y., Wang, D. L. L., Johnson, E. M., and Healy, E. W. (**2018**). "A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions," J. Acoust. Soc. Am. **144**, 1627–1637.