Deep learning based speaker separation and dereverberation can generalize across different languages to improve intelligibility

Eric W. Healy,[1,a)] Eric M. Johnson,[1,b)] Masood Delfarah,[2,c)] Divya S. Krishnagiri,[1,b)] Victoria A. Sevich,[1,b)] Hassan Taherian,[2] and DeLiang Wang[2,b)]

[1]*Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210, USA*

[2]*Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA*

ABSTRACT:

The practical efficacy of deep learning based speaker separation and/or dereverberation hinges on its ability to generalize to conditions not employed during neural network training. The current study was designed to assess the ability to generalize across extremely different training versus test environments. Training and testing were performed using different languages having no known common ancestry and correspondingly large linguistic differences—English for training and Mandarin for testing. Additional generalizations included untrained speech corpus/recording channel, target-to-interferer energy ratios, reverberation room impulse responses, and test talkers. A deep computational auditory scene analysis algorithm, employing complex time-frequency masking to estimate both magnitude and phase, was used to segregate two concurrent talkers and simultaneously remove large amounts of room reverberation to increase the intelligibility of a target talker. Significant intelligibility improvements were observed for the normal-hearing listeners in every condition. Benefit averaged 43.5% points across conditions and was comparable to that obtained when training and testing were performed both in English. Benefit is projected to be considerably larger for individuals with hearing impairment. It is concluded that a properly designed and trained deep speaker separation/dereverberation network can be capable of generalization across vastly different acoustic environments that include different languages. © *2021 Acoustical Society of America*.

https://doi.org/10.1121/10.0006565

(Received 19 February 2021; revised 16 September 2021; accepted 16 September 2021; published online 8 October 2021)

[Editor: G. Christopher Stecker] Pages: 2526–2538

I. INTRODUCTION

Advances in deep learning have substantially elevated our ability to improve speech understanding in complex listening situations. This is of particular importance for the nearly half of a billion people worldwide with hearing loss (World Health Organization, 2020) who typically display poor speech understanding when interfering sounds and/or reverberation are present (e.g., Souza, 2016). Prior to current deep-learning solutions, the problem was particularly intractable, as a single-microphone (speech and interfering sounds picked up by the same microphone) solution capable of improving speech understanding had proven elusive.

The practical efficacy of deep learning based speaker separation and/or dereverberation, as well as noise reduction, hinges on its ability to generalize to conditions not employed during network training. Generalization forms the focus of the current work. The challenge arises because it is obviously impossible to train a neural network in every condition it will encounter, and an improperly trained network can overfit its training environment and display performance limited to that acoustic environment.

This concept of generalization to untrained environments is multifaceted and encompasses a wide variety of differences between training and test environments. In each of the generalization examples that follow, success is dictated by improved intelligibility of target speech following processing to remove interference. Different speech utterances are always used for training and test, and so prior demonstrations possess this generalization (e.g., Healy *et al.*, 2013). Generalization to signal-to-noise or target-to-interferer energy ratios not used during training has also been demonstrated (e.g., Chen *et al.*, 2016). Generalization to untrained background noises represents a particular challenge. However, success has been demonstrated using different segments of the same noise type for training and testing (e.g., Healy *et al.*, 2015; Monaghan *et al.*, 2017; Zhao *et al.*, 2018; Keshavarzi *et al.*, 2019) as well as entirely novel noise types (Chen *et al.*, 2016; Healy *et al.*, 2021). Reverberation characteristics (room impulse responses) have differed across training and test and have been shown to generalize well (e.g., Zhao *et al.*, 2018; Healy *et al.*, 2019; Healy *et al.*, 2020). It has also been possible to

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

[b)]Also at: Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, USA.

[c)]Currently at Apple, One Apple Park Way, Cupertino, CA 95014, USA.

develop talker-independent systems that allow the talker producing the test speech to be absent during training (e.g., Chen and Wang, 2017; Goehring *et al.*, 2017; Goehring *et al.*, 2019; Keshavarzi *et al.*, 2019; Healy *et al.*, 2020).

Other aspects of generalization are perhaps underappreciated. One involves the speech corpora used for training and test (cross-corpus generalization or corpus independence; see Healy *et al.*, 2020; Pandey and Wang, 2020). Differences between corpora involving syllables versus word lists versus sentences are generally well recognized. Sentences are preferred, as they possess greater ecological validity and provide the large signal durations required for network training. However, even corpora all composed of sentences can differ. Sentence lists versus connected speech (e.g., read-aloud passages from written sources) versus spontaneously produced speech can possess different linguistic and acoustic characteristics. These characteristics include co-articulation between sentences, phonetic composition, lexical content, syntactic structure, prosodic structure, semantic predictability, and clarity of articulation (e.g., Underhill, 2005).

A closely related, and perhaps similarly underappreciated, aspect of generalization involves the recording channel. The recording equipment employed to produce the speech materials used for training and test can impart specific characteristics. Notably, the frequency response of the microphone shapes the long-term average spectrum and impulse response of the signal. The acoustic environment in which the speech productions are made and recorded also imparts specific characteristics. Notably, background noise adds to the signal, and the frequency and impulse response/reverberation characteristics of the room modify the signal. Channel characteristics such as these can challenge generalization, particularly at lower signal-to-noise ratios (SNRs), if they differ across training and test utterances. Channel characteristics tend to be constant throughout a given recording, but different across different recordings, causing generalization across channel and the generalization across corpora described just above to be highly related. This issue was examined in detail by Pandey and Wang (2020), who attributed the challenge of cross-corpus generalization largely to differences in recording channels and who propose techniques to improve cross-corpus generalization.

In the current study, a perhaps ultimate generalization is considered—that to an entirely unrelated language. It might be reasonable to assume that a deep learning network could generalize to an untrained language if that language is in the same family and therefore possesses many commonalities. An example might include training on Spanish-language speech materials and testing on Portuguese materials. These languages share a common ancestry, as reflected in their proximity to one another on the language-family trees common to historical linguistics. Accordingly, they share many common phonetic, morphological, lexical, syntactic, and other linguistic characteristics.

In contrast, the greatest degree of generalization was forced in the current study by selecting two widely spoken languages having no known common ancestry—English and Mandarin Chinese. English is a Germanic language with Indo-European roots. Mandarin is a Sinitic language with Sino-Tibetan roots. These entirely independent ancestries cause the languages to be very different in a variety of ways. First, English and Mandarin have different phonetic inventories. English has more vowels, and only two vowels are common across the languages. Mandarin has none of the voiced fricatives of English but has fricatives that English lacks. Mandarin has none of the English voiced stops and only one of the two English approximants. Perhaps related, the orthographic systems are entirely different across these languages. Whereas English uses a Latinate alphabet, in which letters correspond roughly to individual phonemes, Mandarin uses a logographic system in which symbols represent words. Orthography can potentially influence phonological representations and production of phonemes (e.g., Ranbom and Connine, 2007). For example, the commonly produced English flap [ɾ] is often represented as the written letter ⟨t⟩, and that orthographic representation can potentially shift pronunciation toward the voiceless alveolar stop [t] in some situations. English and Mandarin also differ in their prosody and the rhythm with which they are spoken and perceived. Whereas English is a stress-timed language, Mandarin is a syllable-timed language. Finally, Mandarin is a tonal language whereas English is not. Accordingly, pitch contours of the voice carry important semantic cues in Mandarin, whereas pitch contours in English primarily code prosodic information.

The current study was designed to establish the ability of a deep learning speech-processing algorithm to generalize across vastly different training versus test environments. Training was performed using English-language speech materials, and testing was performed using Mandarin-language speech materials. The algorithm was required to perform separation of two concurrent talkers and dereverberation to increase intelligibility of a target talker. A deep Computational Auditory Scene Analysis (deep CASA) algorithm was employed. In addition to cross-language generalization, the algorithm was tasked with generalizing across corpora and channel, to untrained target-to-interferer intensity ratios (TIRs), and to untrained room impulse responses (RIRs). Finally, the network was talker independent and evaluated using a talker not involved in training.

II. METHOD

A. Subjects

Ten native speakers of Mandarin Chinese were recruited from The Ohio State University and surrounding community. All had normal hearing, as indicated by pure-tone audiometric thresholds of 20 dB HL or lower at octave frequencies from 250 to 8000 Hz on day of test (ANSI, 2004, 2010a). Ages ranged from 23 to 46 years (mean = 30.2), four were female, and six were male. None had any prior experience with the test-sentence materials used in the current study, and all received a monetary incentive for participating.

B. Stimuli

The stimuli used for training and test consisted of sentence pairs, one target and one interfering sentence, both having substantial amounts of room reverberation. All training and test signals were processed at 16 kHz sampling with 16 bit resolution.

The stimuli used for training the deep learning algorithm involved English-language speech from the Wall Street Journal Continuous Speech Recognition Corpus (WSJ0; Paul and Baker, 1992). This corpus contains recordings from a large number of talkers reading Wall Street Journal newspaper articles from the late 1980s (approximately 39 000 sentences, 80 h of audio). Accordingly, it represents continuous speech. It was designed to support the development of automatic speech recognition systems and is commonly used for training and testing talker-independent speaker separation (Hershey *et al.*, 2016). Sentences were drawn from the si_tr_s folders, which contain recordings from 101 talkers (49 male and 52 female), each of whom produced an average of 124 sentences.

Training-sentence pairs were created by selecting sentences produced by different WSJ0 talkers and equating them to the same root mean square level. This equating produced a single training TIR of 0 dB, different from the TIRs used for testing. The talkers comprising each sentence pair were always different and either could be male or female. The longer-duration sentence of each pair was trimmed to match the duration of the shorter, to avoid periods containing a single talker.

To generate room reverberation, a 6 m×7 m×3 m simulated room was used. The room reverberation time (T_{60} value) for each sentence pair was selected randomly from 0.3 to 1.0 s. The target talker was placed 1 m from the microphone position (the listener), and the interfering talker was placed 2 m away, both at the same elevation as the microphone. The microphone position was fixed in the room at (3, 4, 1.5) m, and each talker was positioned at one of 36 randomly selected angles evenly distributed around the microphone. Each individual sentence of each pair was convolved with an RIR to generate reverberant speech, using an RIR generator (Habets, 2020) that implements the image method (Allen and Berkley, 1979).

The reverberant sentence recordings for each pair were mixed to generate the reverberant two-talker mixtures. A total of 200 500 training mixtures were generated, from which 500 were reserved for cross-validation. These training utterances and their processing were identical to those employed by Healy *et al.* (2020), who performed training and testing both in the same language, in order to facilitate direct comparison to the current cross-language conditions.

The sentence pairs used for testing (inference) were created using different corpora containing Mandarin speech materials. These were produced by talkers different from those producing the training materials. The target test sentences were drawn from the Mandarin Speech Perception test (MSP, Fu *et al.*, 2011), and the interfering test sentences were drawn from the Tsinghua Chinese 30 h database (THCHS-30, Wang and Zhang, 2015).[1]

The MSP contains 100 sentences arranged into ten lists. Each sentence contains seven monosyllabic words selected to be familiar and widely used in daily life. The sentences are phonetically balanced in terms of vowels, consonants, and tones, with proportions approximating those of natural everyday Mandarin speech. The standard recording of these materials was used, which was produced by a female professional radio broadcaster at a natural speaking rate. Fu *et al.* (2011) reported 100% recognition accuracy in quiet for normal-hearing (NH) speakers of Mandarin. For the current study, 80 sentences were used for testing and 20 were reserved for practice.

The THCHS-30 includes sentences read from an encyclopedic reference book and, like the WSJ0, was designed to support the development of automatic speech recognition systems. The corpus contains approximately 35 h of speech produced by 40 native speakers of Mandarin. Productions from male talker D8 were selected for the current study. Of the 250 sentences produced by this talker, 100 sentences were selected, with each approximating the duration of a paired sentence from the MSP.

The sentences from the two Mandarin test corpora were arranged into pairs, with the interfering sentence always longer in duration than the target sentence. The use of different-gender talkers for these sentence pairs allowed the human subjects to differentiate and identify which talker was the target. It has been shown that deep learning algorithm performance is similar when target and interfering talkers are different genders versus when they are both male or both female (see Table V from Liu and Wang, 2019).

The preparation of the test-sentence pairs followed that for the training-sentence pairs. Exceptions were that each sentence pair was mixed at the two test TIRs of –8 and –5 dB, chosen to produce intelligibility values generally free of floor and ceiling effects. With regard to room reverberation, each of the talker positions in the virtual room was shifted by five degrees relative to the training positions, to produce different RIRs across training and test. The test T_{60} values were 0.6 and 0.9 s, and each test-sentence pair was prepared in each of these TIR/T_{60} conditions.

C. Deep learning algorithm description

The algorithm was a variant of deep CASA, which was proposed for talker-independent speaker separation by Liu and Wang (2019), then extended to reverberant conditions by Healy *et al.* (2020). Figure 1 displays its basic function. Deep CASA contains a simultaneous grouping stage followed by a sequential grouping stage. Simultaneous grouping involves the separation of acoustic components from the two sound sources in each time frame. Dereverberation is also performed during this stage. Sequential grouping involves the organization of the separated frames into two streams, one for each talker. The rationale for such a two-stage model comes from classic theories of auditory scene analysis, in which the human processing of sound is hypothesized to involve such stages (Bregman, 1990), and from computational auditory scene analysis (Wang and Brown, 2006), which often employs such stages. In the current study, the first stage was accomplished using a U-Net convolutional neural network with densely connected layers (Dense-UNet; Liu and Wang, 2019), and the second stage employed a temporal convolutional network (TCN; Bai *et al.*, 2018; Lea *et al.*, 2016). Note that deep CASA is a dedicated speaker separation algorithm. The network also performed de-reverberation, but noise removal was not addressed.

In the current study, both the amplitude and phase of the signal of interest were obtained by working in the complex domain. In this approach, the real and imaginary parts are both estimated by the deep neural network, which allows both the amplitude and phase of the signal of interest to be obtained. This approach contrasts with that of most studies, in which only the amplitude representation of the signal of interest is estimated by the deep neural network, then combined with the phase of the original unprocessed sound mixture ("noisy phase") to reconstruct the isolated signal of interest. The current training target was the complex ideal ratio mask (cIRM, Williamson *et al.*, 2016), and the features involved the real and imaginary components of the complex short-time Fourier transform (STFT).

The current algorithm and its training were identical to that employed by Healy *et al.* (2020). In that study, talker-independent reverberant speaker separation was performed, but the network was trained and tested both using English-language speech materials. The use of an identical algorithm allowed the current cross-language results to be directly compared to the previous within-language results. Accordingly, the interested reader is directed to Healy *et al.* (2020) and Liu and Wang (2019) for additional details on the deep neural network.

The current model was non-causal. This is in accord with our overall approach, in which we first establish high performance benchmarks through the use of unconstrained networks. We then address implementation concerns, including causal operation. This two-stage approach allows the ramifications of each implementation modification to be known.

The mixture signal $y(t)$ can be expressed as,

$$y(t) = h_1(t) * s_1(t) + h_2(t) * s_2(t), \tag{1}$$

where $s_1(t)$ and $s_2(t)$ are the two anechoic talker signals, $h_1(t)$ and $h_2(t)$ are the RIRs corresponding to each speaker location in the room, and $*$ denotes convolution. The computational problem is defined as extracting $s_1(t)$ and $s_2(t)$ from $y(t)$.

1. Simultaneous grouping

Dense-Unet (Liu and Wang, 2019) extends the U-net architecture (Ronneberger *et al.*, 2015) by interleaving

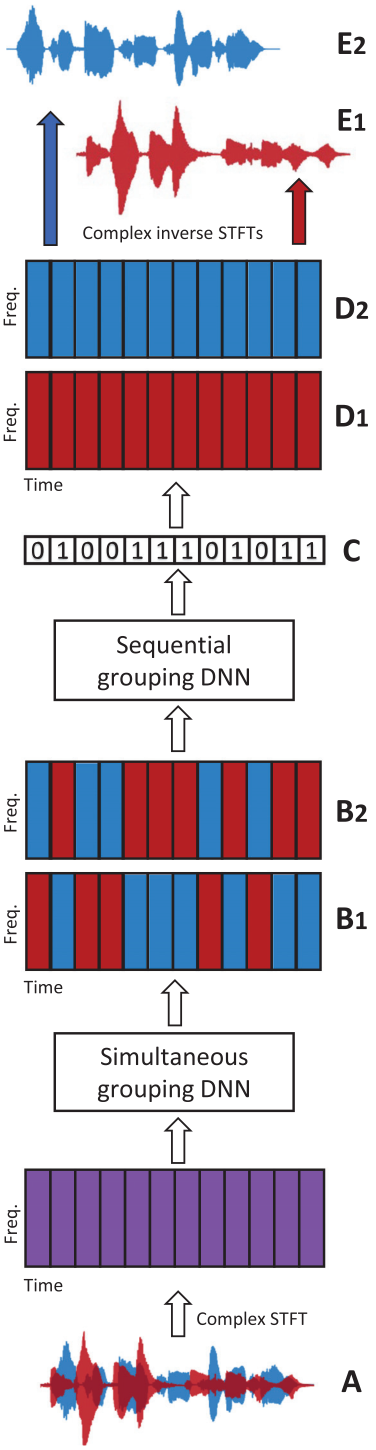

FIG. 1. (Color online) A schematic of the current deep CASA algorithm to separate and dereverberate simultaneous talkers. (A) is the reverberant two-talker mixture signal; (B_1) and (B_2) are the signals corresponding to the two talkers dereverberated and separated in each time frame, without regard to which talker is which; (C) is the predicted frame-talker assignment vector, where 0 indicates that the talker assignment in stage B is correct and 1 indicates that the talker assignment needs to be reversed in that frame; (D_1) and (D_2) are the separated and dereverberated signals corresponding to the two individual talkers; and (E_1) and (E_2) are the corresponding output waveforms. The simultaneous grouping DNN was a U-Net convolutional neural network with densely-connected layers, and the sequential grouping DNN was a temporal convolutional network. Both the amplitude and phase of the signal of interest were obtained by working in the complex domain.

dense blocks between its layers. Dense blocks were originally introduced with the DenseNet architecture (Huang *et al.*, 2017). In the current Dense-Unet implementation, a U-net was used having four upsampling and four downsampling layers, with each layer interleaved by a dense block.

The input to this simultaneous grouping network was real and imaginary STFT features $M(m,f)$, where m represents the frame index and f is the frequency channel. This network was used to estimate two cIRMs, one for each talker. Pointwise multiplying these masks by $M(m,f)$ in the complex domain resulted in two STFT signals $\hat{S}_{u_1}(m,f)$ and $\hat{S}_{u_2}(m,f)$ that represent the separated and dereverberated frames for each talker.

Frame-level permutation invariant training (tPIT; Kolbaek *et al.*, 2017) was used as the training loss. Accordingly, two loss functions l_1 and l_2 were calculated for each time frame,

$$l_1(m) = \Sigma_f \left|\hat{S}_{u_1}(m,f) - S_1(m,f)\right| + \Sigma_f \left|\hat{S}_{u_2}(m,f) - S_2(m,f)\right|, \quad (2)$$

$$l_2(m) = \Sigma_f \left|\hat{S}_{u_1}(m,f) - S_2(m,f)\right| + \Sigma_f \left|\hat{S}_{u_2}(m,f) - S_1(m,f)\right|, \quad (3)$$

where $S_1(m,f)$ and $S_2(m,f)$ are the clean anechoic talker signals.

Next, frames were assigned to each talker based on the smaller loss. These optimally organized STFT features were converted to time-domain signals, from which the SNR-based loss function J^{SNR} was calculated and minimized,

$$J^{SNR} = -10 \sum_{i=1,2} \log \frac{\sum_t s_i(t)^2}{\sum_t \left[s_i(t) - \hat{s}_{o_i}(t)\right]^2}, \quad (4)$$

in which $s_i(t)$ are the clean anechoic signals and $\hat{s}_{o_i}(t)$ are their corresponding estimations.

2. Sequential grouping

The TCN is able to capture long-range contextual information due to its series of dilated convolutions and resulting large receptive fields. This is desirable for speech processing and enables the tracking of a talker over a long utterance. The current TCN had eight dilated convolutional blocks, each with three convolutional layers.

The previous simultaneous stage was trained using optimal talker-frame assignments based on the signals $s_1(t)$ and $s_2(t)$, which are not available at test time. Therefore, the current sequential grouping stage used the outputs of the first stage $\hat{S}_{u_1}(m,f)$ and $\hat{S}_{u_2}(m,f)$ and was trained to predict a temporal organization vector $\boldsymbol{A}$ per frame. Specifically, $\boldsymbol{A} = [1,0]$ indicated that $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$ correctly represent frames for talker 1 and talker 2, respectively, whereas $\boldsymbol{A} = [0,1]$ meant that the frames need to be reversed. Separated frames $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$ will be optimally organized over the sentence if $\boldsymbol{A}$ is optimally predicted. To predict $\boldsymbol{A}$, the network generated an embedding vector $\boldsymbol{V}(c) \in \mathbb{R}^d$ for each time frame c, where d is the size of the embedding vector, which was then optimized with the loss function (Hershey *et al.*, 2016; Liu and Wang, 2019),

$$J^{DC} = \|\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{A}\boldsymbol{A}^T\|_F^2, \quad (5)$$

in which $\|.\|_F$ denotes the Frobenius norm.

At inference time, the network generated the embeddings $\boldsymbol{V}(c)$, and a K-means algorithm clustered these vectors into two groups, labeled as $\hat{A}(m) = \{0,1\}$, which were used to organize $\hat{S}_{u_1}(m,f)$ and $\hat{S}_{u_2}(m,f)$. Finally, these spectrograms were converted to time-domain signals $\hat{s}_1(t)$ and $\hat{s}_2(t)$ *via* inverse STFT, which are the estimated anechoic talker signals. The simultaneous and sequential grouping stages were trained separately, and the training was stopped when no further improvement in the cross-validation set was achieved.

Figure 2 displays spectrogram images for an example stimulus. Figure 2(a) displays two Mandarin sentences mixed at –5 dB TIR and reverberated using a T_{60} value of 0.9 s. Figures 2(b) and 2(c) display the individual sentences prior to mixing and reverberation. Figures 2(d) and 2(e) display these Mandarin sentences separated from the reverberant two-talker mixture, Fig. 2(a), using the current deep CASA algorithm, trained using English-language speech materials.

D. Procedure

Listeners heard sentence pairs in unprocessed conditions (pairs of sentences mixed and reverberated) and in processed conditions (pairs of reverberant sentences processed by the deep CASA algorithm to isolate the target talker and remove reverberation). There were a total of eight conditions heard by each listener (2 unprocessed/processed × 2 TIRs × 2 T_{60}s). Each listener heard 80 sentence pairs, with ten sentences in each condition. The critical comparison is between unprocessed and processed in each condition, so these unprocessed/processed conditions were heard in juxtaposed order within each TIR-T_{60} block. The order of the four TIR-T_{60} blocks was randomized for each listener, and the order of unprocessed/processed was randomized for each listener in each TIR-T_{60} block. The sentence materials were presented in a single fixed order for all listeners to yield a random correspondence between sentence pairs and condition for each listener.

The stimuli were played back from a Windows PC using an RME Fireface UCX digital-to-analog converter (Haimhausen, Germany), routed through a Mackie 1202-VLZ mixer (Woodinville, WA), and delivered diotically using Sennheiser HD 280 Pro headphones (Wedemark, Germany). Each stimulus was scaled to the same total root mean square level and set to play back at 65 dBA in each ear, as measured by an ANSI type I sound-level meter and flat-plate headphone coupler (Larson Davis models 824 and

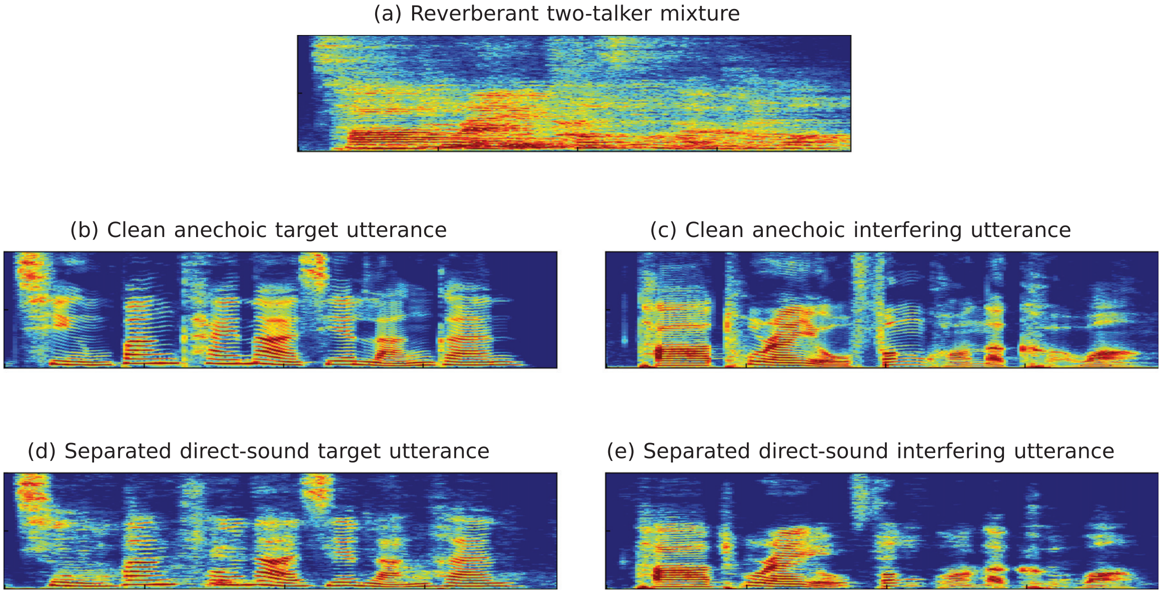

FIG. 2. (Color online) Panel (a) displays a spectrogram image of two competing Mandarin-language sentences (target and interferer) having a large amount of room reverberation ($T_{60} = 0.9$ s). Panels (b) and (c) display the individual non-reverberant sentences prior to mixing. Panels (d) and (e) display the target and interfering Mandarin sentences dereverberated and extracted from the mixture (a) by the deep CASA algorithm, which was trained using English-language speech. Compare the output (d) to the desired signal (b) and the output (e) to the desired signal (c).

AEC 101, Depew, NY). Signal calibration was conducted prior to testing each listener.

In a brief familiarization immediately preceding formal testing, listeners heard the 20 practice sentences arranged into five blocks, with four sentence pairs (or clean MSP sentences) in each block. The practice conditions were: (1) interference-free, reverberation-free speech spoken by the target talker, (2) algorithm processed sentence pairs in the most favorable TIR-T_{60} condition (–5 dB, 0.6 s), (3) algorithm processed sentence pairs in the least favorable condition (–8 dB, 0.9 s), (4) unprocessed sentence pairs in the most favorable TIR-T_{60} condition, and (5) unprocessed sentence pairs in the least favorable TIR-T_{60} condition. These sentence pairs were distinct from those used for testing. Listeners were instructed to attend to the female voice, repeat back each sentence as best they could, and to guess if unsure of what was said.

Following familiarization, listeners heard the eight blocks of experimental conditions, receiving the same instructions as for practice. They were seated alone in a double-walled audiometric booth. The experimenter controlled the presentation of each stimulus from a position just outside the booth, with the listener in view through a large window. Each stimulus was presented only once to each listener, and no feedback was provided during testing.

Listener responses were recorded digitally using a Shure SM11 microphone (Niles, IL) positioned inside the audiometric booth. These responses were scored off-line by two native speakers of Mandarin, who were blind to the condition under test and to one-another's scoring. Scorers listened to each recorded response as many times as needed and documented how many words were correctly reported for each target sentence.

III. RESULTS AND DISCUSSION

A. Human performance

Inter-rater reliability between the two scorers was assessed using a two-way mixed, absolute agreement, average-measures interclass correlation (ICC; McGraw and Wong, 1996) on rationalized arcsine units (RAUs, Studebaker, 1985). The resulting ICC of 0.997 was in the excellent range (Cicchetti, 1994), indicating that the two scorers had a high degree of agreement. The excellent ICC suggests that the statistical power of subsequent analyses was not substantially reduced by measurement error. Accordingly, sentence intelligibility was defined as the percentage of words correctly reported by each listener in each condition, using values averaged across the two scorers. The seven words per sentence and ten sentences per condition yielded a total of 70 words in each condition for each listener.

Figure 3 displays intelligibility for each individual listener in each condition: The two TIRs employed are plotted in separate panels. In each panel, the unprocessed and processed conditions for a given T_{60} are represented by adjacent columns. Each T_{60} is grouped into a pair of columns for each listener, with 0.9 s on the left (white solid and white hatched columns) and 0.6 s on the right (black solid and gray hatched columns). Algorithm benefit for each listener in a given condition corresponds to the difference between a solid column (unprocessed) and the hatched column directly to the right (processed). Note that NH9 was unable to correctly report any words in one condition, so that column is absent (TIR = –8 dB, $T_{60} = 0.6$ s, unprocessed).

As Fig. 3 shows, every listener received algorithm benefit in every condition, except for one case where the

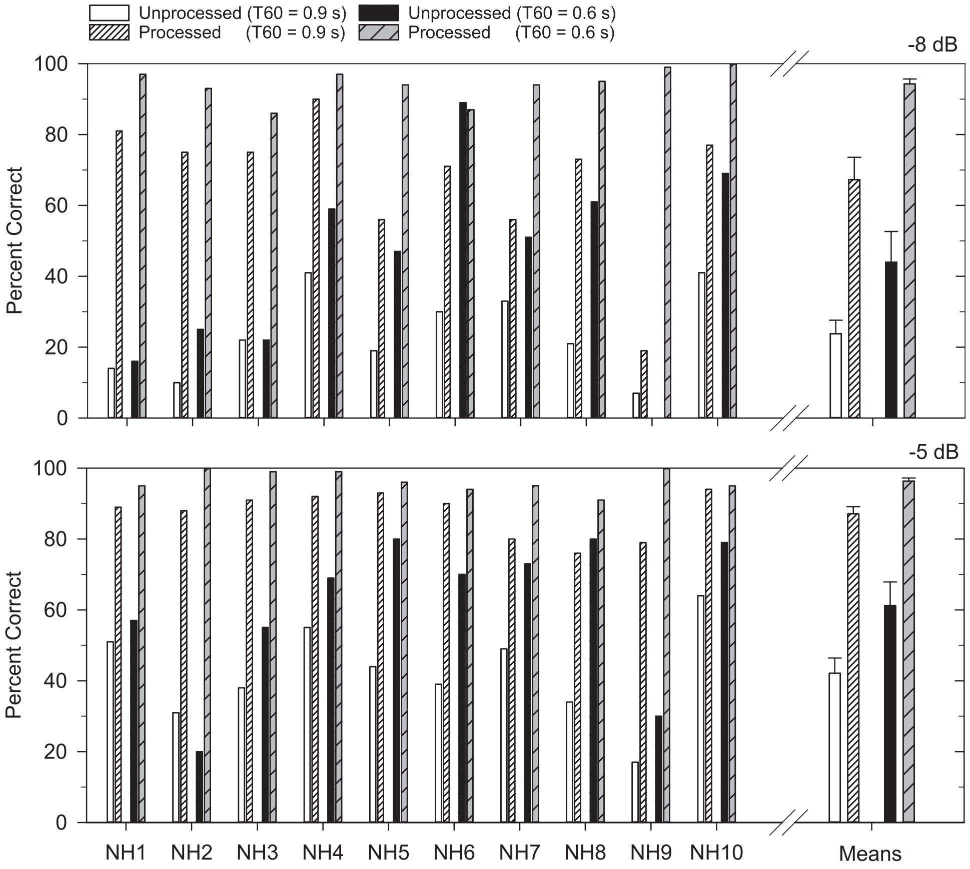

FIG. 3. Intelligibility of Mandarin sentences in each condition. Shown are scores for individual NH listeners as well as group means (and SEs). Unprocessed scores result from the mixture of a target and interfering Mandarin sentence, both having a substantial amount of room reverberation. Processed scores result from this signal following processing by the deep CASA algorithm, trained using English-language speech, to dereverberate and isolate the target talker. The target-to-interferer ratios of –8 and –5 dB are displayed in separate panels, and the T_{60} times of 0.9 and 0.6 s are displayed in each panel using different columns. Algorithm benefit is obtained by comparing each unprocessed column (unhatched) to the immediately adjacent processed column (hatched).

unprocessed score was already high at 88.6% correct (NH6, TIR = –8 dB, T_{60} = 0.6 s). Algorithm benefit is in part a function of unprocessed score, and accordingly, was correlated with unprocessed scores ($|r| = 0.75$, $p < 0.0001$, across all conditions), where lower unprocessed scores tended to be associated with greater benefit. Across all listeners and conditions (40 cases), benefit was 20% points or greater in 88% of cases, 40% points or greater in 53% of cases, and 60% points or greater in 23% of cases.

Figure 3 also displays group-mean intelligibility scores and standard errors (SEs) for each condition. The top panel again represents the less favorable TIR of –8 dB. At this TIR, the less favorable T_{60} produced the lowest mean unprocessed score (23.8% correct) and a mean benefit of 43.5% points. The more favorable T_{60} at this TIR produced a mean unprocessed score of 44.0% correct and the largest mean benefit of any condition (50.3% points). At the more favorable TIR of –5 dB (bottom panel), the less favorable T_{60} produced a mean unprocessed score of 42.1% correct and a benefit of 45.0% points. The more favorable T_{60} at this TIR produced the highest group-mean unprocessed score (61.2% correct) and the correspondingly smallest algorithm benefit (35.1% points). The grand-mean algorithm benefit across conditions was 43.5% points.

It is notable that group-mean algorithm-processed scores approached the ceiling at 100% correct in both of the more favorable T_{60} conditions (94.3% and 96.4% correct, see Fig. 3 rightmost column in each panel). Because near-ceiling intelligibility values were observed, benefit was examined in RAUs, which counteract the compression of percent correct against the floor or ceiling. The RAU benefits for the four conditions were as follows: TIR –8 dB, T_{60} 0.9 s = 43.2 points; TIR –8 dB, T_{60} 0.6 s = 58.7 points; TIR –5 dB, T_{60} 0.9 s = 46.5 points; TIR –5 dB, T_{60} 0.6 s = 44.5 points. The grand-mean algorithm benefit across conditions was 48.2 RAU points.

The primary analysis consisted of planned comparisons, which were uncorrected, two-sided, paired, *t*-tests on RAUs, performed to examine algorithm benefit in each condition. Scores were treated as independent samples when calculating effect sizes (Cohen's *d*). Scores for algorithm-processed conditions were significantly higher than the corresponding unprocessed scores in all four conditions:

TIR -8 dB, T_{60} 0.9 s: $t(9) = 8.1$, $p < 0.0001$, Cohen's $d = 2.54$;
TIR -8 dB, T_{60} 0.6 s: $t(9) = 5.5$, $p < 0.001$, Cohen's $d = 2.58$;
TIR -5 dB, T_{60} 0.9 s: $t(9) = 14.5$, $p < 0.000001$, Cohen's $d = 4.23$;
TIR -5 dB, T_{60} 0.6 s: $t(9) = 5.1$, $p < 0.001$, Cohen's $d = 2.87$.

These results all survive Bonferroni correction for multiple comparisons.

A supplementary statistical analysis was performed using a linear mixed-effects model. The outcome variable was the RAU-transformed percent-correct scores for each listener in each condition (80 data points). The fixed effects were processing condition, TIR, and T_{60}, in addition to each of the two- and three-way interactions between the first-order effects. The model also included random intercepts for listener. Deviation coding was used to represent the variables of processing condition (unprocessed coded as -0.5, processed as 0.5), TIR (–8 dB coded as -0.5, –5 dB as 0.5), and T_{60} (0.9 s coded as -0.5, 0.6 s as 0.5). Visual inspection of residual plots did not reveal any apparent violations of homoscedasticity or normality. The analysis was performed using R 4.0.3 (R Core Team, 2020) and the lme4 package (Bates *et al.*, 2015). Degrees of freedom for the t distribution (two-sided) were based on Satterthwaite's approximation using the lmerTest package (Satterthwaite, 1941; Kuznetsova *et al.*, 2020).

Most notably, the fixed effect of algorithm processing was large and significant, reflecting that the algorithm was successful in increasing listeners' overall intelligibility [$\beta = 48.4$ RAU points, $SE = 3.27$, $t(63) = 14.8$, $p < 0.0001$]. The fixed effect of TIR was also significant, simply reflecting overall higher intelligibility at the higher (more favorable) TIR [$\beta = 15.6$ RAU points, $SE = 3.27$, $t(63) = 4.8$, $p < 0.0001$]. Finally, the fixed effect of T_{60} was also significant, simply reflecting overall higher intelligibility at the lower (more favorable) reverberation time [$\beta = 21.8$ RAU points, $SE = 3.27$, $t(63) = 6.7$, $p < 0.0001$]. None of the two- or three-way interactions were significant, indicating no significant interdependency between any of these effects [each $t(63) < 1.4$, each $p > 0.15$].[2]

Figure 4 displays group-mean intelligibility for the Mandarin sentences observed currently, along with the intelligibility of English-language sentences from Healy *et al.* (2020), in the common conditions of TIR = –5 dB. Across studies, the preparation of two-talker reverberant sentences was identical (except that different speech corpora were used), and the test apparatus and procedures were largely identical. Also, recall that the design and training of the deep CASA algorithm using English-language speech was identical across studies. The only notable difference across studies involved the speech materials used for testing—Mandarin currently and English by Healy *et al.* (2020). Accordingly, the current study involved cross-language generalization whereas the former did not. Intelligibility for the T_{60} value of 0.9 s is displayed on the left half of the figure whereas scores for the T_{60} value of 0.6 s are on the right. As in the previous figures, unprocessed scores are represented by solid columns, whereas processed scores are represented by hatched columns, allowing benefit to be determined as the difference between each pair of adjacent columns. The labels at the top of the figure distinguish the current data ("Language: Across"—trained on English/tested on Mandarin) from those of the previous study ("Language: Within"—trained and tested on English).

As can be seen in Fig. 4, the benefit observed currently is comparable to that observed previously, despite the current addition of generalization to a language having no shared ancestry. Averaged across the two T_{60} conditions, the difference in benefit across studies is 6.2 raw percentage points, with the current benefit being slightly lower (difference is 2.4% points at $T_{60} = 0.9$ s and 10.1% points at $T_{60} = 0.6$ s). When intelligibility is converted to RAUs, which is needed to account for the one near-ceiling mean value in Fig. 4, the overall average benefit difference across studies becomes 1.7 points (difference remains similar at 2.0 points at $T_{60} = 0.9$ s, but is reduced to 1.3 points at $T_{60} = 0.6$ s, where the ceiling algorithm-processed mean of 96.4% is observed).

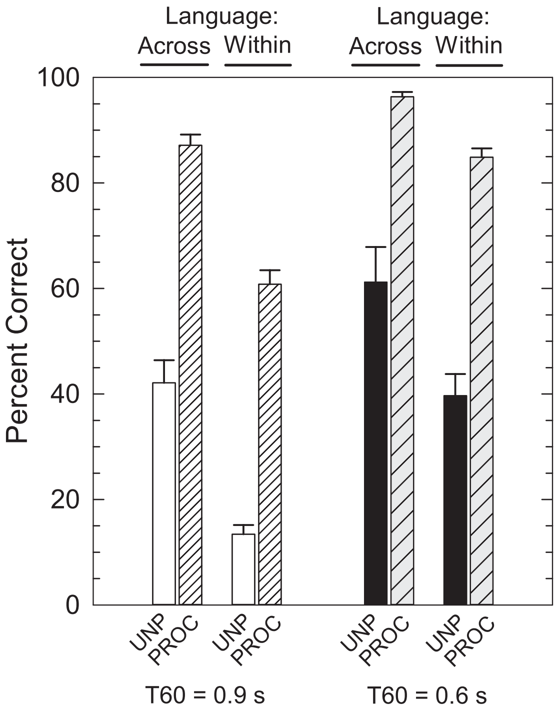

FIG. 4. A comparison between across-language and within-language algorithm performance. Displayed are group-mean sentence intelligibilities (and SEs) in unprocessed and algorithm-processed conditions at a TIR of –5 dB. The two reverberation T_{60} times are on the left and right sides of the panel. The columns labeled "Language: Across" are from the current Fig. 3, where training was performed using English-language speech materials, and testing was performed using Mandarin-language speech materials. The columns labeled "Language: Within" are from Healy *et al.* (2020), who employed an identical algorithm and identical training stimuli but instead tested using English-language speech materials.

This across-language versus within-language comparison was assessed using a linear mixed-effects model. The outcome variable was the RAU-transformed percent-correct score for each listener (ten native Mandarin, ten native English), at the common TIR of –5 dB, in two algorithm-processing conditions, at two T_{60}s, amounting to 80 scores. The fixed effects for this model were processing condition, T_{60}, and language, as well as each of the two- and three-way interactions. Again, the model included random intercepts for listener ($n = 20$), and deviation coding was used for each independent variable, with the baseline condition coded as -0.5 and the comparison condition coded as 0.5. The unprocessed algorithm condition was coded as the baseline, and the algorithm-processed condition was coded as the comparison. The less favorable TIR of 0.9 s was coded as the baseline, and the more favorable TIR of 0.6 s was coded as the comparison. English was arbitrarily assigned as the baseline language, and Mandarin was designated as the comparison language. No violations of homoscedasticity or normality were apparent in the residual plots. The analysis was performed using the same software and Satterthwaite's approximation described above.

The important comparison involves the benefit of algorithm processing in across- versus within-language conditions. No significant interaction was observed between algorithm processing and language [β = –1.6 RAU points, SE = 4.82, $t(54)$ = –0.33, p = 0.74], thus providing no evidence that the algorithm was significantly more effective at improving intelligibility when trained in English and tested in Mandarin versus when trained and tested both in English. The large and significant effect of processing [β = 46.4 RAU points, SE = 2.41, $t(54)$ = 19.3, $p < 0.0001$] reflected that the algorithm increased intelligibility across T_{60}s and languages. The large and significant effect of T_{60} [β = 22.2 RAU points, SE = 2.41, $t(54)$ = 9.2, $p < 0.0001$] simply reflected the detrimental effect of increased reverberation on intelligibility. This detrimental effect was stronger in English than in Mandarin, as indicated by a significant interaction between T_{60} and language [β = –9.7 RAU points, SE=4.82, $t(54)$ = –2.0, p = 0.05]. Intelligibility was also higher overall for the Mandarin-language target talker than the English-language target talker, as indicated by the significant effect of language [β = 25.0 RAU points, SE = 3.04, $t(18)$ = 8.2, $p < 0.0001$]. This effect is likely due to the Chinese talker's experience as a professional radio presenter and the tendency for these individuals to produce clear speech. Also, the MSP test employed common everyday vocabulary, whereas the English utterances were less commonplace (see Sec. IV). Neither the interaction between processing condition and T_{60} nor the three-way interaction were significant [each $|\beta| < 2.3$ RAU points, each $SE > 4.8$, each $|t(54)| < 0.47$, each $p > 0.64$].

B. Objective measures

Objective measures are based on measurement of the acoustic stimuli themselves and can be beneficial for comparing performance across studies involving stimulus processing because the variability associated with human subjects is absent. In the current study, two objective measures of speech intelligibility, one measure of speech quality, and one measure of TIR improvement were calculated. All are commonly employed measures and were based on the 80 Mandarin sentence pairs used for testing.

The objective measures of intelligibility included ESTOI (extended short-time objective intelligibility; Jensen and Taal, 2016), and STOI (short-time objective intelligibility; Taal *et al.*, 2011). Both are essentially correlations between the amplitude envelopes of (a) the original interference-free, reverberation-free speech and (b) the same speech utterance following corruption (here, by interference and reverberation) then algorithm processing to remove the corruption. The use of amplitude envelopes reflects their perceived importance to human speech recognition (e.g., Rosen, 1992; Healy and Warren, 2003; Apoux and Healy, 2013), and higher values indicate that the algorithm output more veridically matches the desired clean speech. Because it is a correlation, the scale typically ranges from 0 to 100% (or 0.0 to 1.0). As Table I shows, benefit (processed minus unprocessed) reflected by ESTOI scores ranged from 32.4% to 44.3% points, with a mean across conditions of 40.0% points. STOI benefit ranged from 24.0% to 33.2% points, with a mean of 29.9% points. These objective values tend to underestimate the actual human-subjects benefit observed currently.

The sound-quality prediction was PESQ (Rix *et al.*, 2001). It also reflects a comparison between clean and processed speech and has a scale ranging from −0.5 to 4.5. Increases in PESQ ranged from 0.8 to 1.4, with a mean benefit of 1.1. Although sound quality was not assessed currently by human listeners, this substantial increase in PESQ score suggests that human listeners should consider the current cross-language algorithm-processed speech to have improved sound quality.

Finally, source-to-distortion improvement (ΔSDR; Vincent *et al.*, 2006) reflects the SNR improvement (TIR currently) in dB resulting from processing. The current values range from 11.0 to 13.0 dB, reflecting the substantial ability of the cross-language algorithm to isolate the target signal.

These objective measures may be compared to those from Table I in Healy *et al.* (2020), which provides a direct comparison to a within-language model. This comparison is presented in Table II, where values are expressed as algorithm benefit. ESTOI benefit was within 1.8% points at both T_{60} values. STOI benefit was within 2.9 and 4.9% points at T_{60} values of 0.6 and 0.9 s. ΔSDR was within 1.5 and 1.8 dB

TABLE I. Average ESTOI, STOI, PESQ, and ΔSDR values in different room reverberation (T_{60}) and target-to-interferer ratio (TIR) conditions for the target talker in reverberant two-talker mixtures prior to and following processing by deep CASA. The deep learning algorithm was trained on English-language speech materials and tested on Mandarin-language speech materials. Comparable values for the same algorithm trained on English-language speech materials and tested on English-language speech materials are available in Healy *et al.* (2020) (see their Table I).

		0.6 T_{60} (s)		0.9 T_{60} (s)		
	TIR (dB)	−8	−5	−8	−5	Average
ESTOI (%)	Unprocessed	20.30	24.60	15.20	17.70	19.45
	Processed	61.50	68.90	47.60	59.70	59.43
	Benefit (% pts)	41.2	44.3	32.4	42.0	40.0
STOI (%)	Unprocessed	44.50	48.50	42.10	44.80	44.98
	Processed	76.30	81.70	66.10	75.20	74.83
	Benefit (% pts)	31.8	33.2	24.0	30.4	29.9
PESQ	Unprocessed	0.35	0.40	0.34	0.38	0.37
	Processed	1.53	1.81	1.16	1.52	1.51
	Benefit	1.2	1.4	0.8	1.1	1.1
ΔSDR (dB)	Proc-Unp	13.00	12.48	11.03	11.65	12.04

TABLE II. Average ESTOI, STOI, and PESQ benefit (processed−unprocessed scores), along with ΔSDR values, for across-language versus within-language conditions. The across-language conditions involved training on English-language speech materials and testing on Mandarin-language speech materials. The within-language conditions involved training on English-language speech materials and testing on English-language speech materials. The T_{60} values of 0.6 and 0.9 s and the target-to-interferer ratio of−5 dB common across studies were considered.

	$T_{60} = 0.6$ s			$T_{60} = 0.9$ s		
	Across language	Within language	Difference (across-within)	Across language	Within language	Difference (across-within)
ESTOI Benefit (% points)	44.30	46.07	−1.8	42.00	43.80	−1.8
STOI Benefit (% points)	33.20	36.10	−2.9	30.40	35.29	−4.9
PESQ Benefit	1.41	1.22	0.2	1.14	1.00	0.1
ΔSDR (dB)	12.48	13.95	−1.5	11.65	13.42	−1.8

at these T_{60} values. In each of these cases, the current cross-language benefits were lower than the previous within-language benefits. With regard to PESQ, benefit values were within 0.2 and 0.1 for the two T_{60} values, with the current cross-language benefits being slightly higher.

IV. GENERAL DISCUSSION

The current study was designed to demonstrate the ability of a well-designed and trained deep learning algorithm to generalize across extremely different acoustic environments. But more direct practical implications may also exist. The languages chosen currently are the world's most common, with over a billion speakers each. English has the largest number of total speakers in the world, with Mandarin close behind, and Mandarin has the world's largest number of native speakers (Eberhard *et al.*, 2020). These are therefore important global languages with a substantial need for speech technology.

Speaker-separation or noise-reduction that operates optimally on a single primary language is still of substantial value. However, the ability of a system to operate across languages, as demonstrated currently, substantially increases its value. This is an important consideration because of the challenge associated with training on each of the world's 7000-plus currently spoken languages, or even on the 200 most widely spoken languages, which represent the native languages of 88% of the world population (Eberhard *et al.*, 2020). The current results suggest that people speaking languages for which resources to collect and train models are lacking might nevertheless benefit directly from the vast speech and language data that have been collected in other languages, such as English and Mandarin. These current results are also relevant to the related issue involving the varied regional dialects that a given language possesses, which even a single-language algorithm would need to be robust to.

The current differences across training and test were vast and represent a highly challenging generalization. Further, the listening conditions were highly complex and challenging, both for a processing algorithm and for NH listeners, with the latter often able to tolerate challenging acoustic conditions. The room reverberation applied to the concurrent talker conditions was substantial. The T_{60} value of 0.6 s corresponds to the upper limit for acceptable reverberation in classrooms (ANSI, 2010b), whereas the value of 0.9 exceeds that limit. Despite these challenges, significant benefit was observed in all conditions. It is also notable that, with one exception (1 of 40 cases), the algorithm did not produce any decrement in performance. This is an important consideration because the possibility exists that the substantial processing could distort the signal and decrease performance when baseline scores are high and benefit is not needed. In the one instance of decrement (NH6, TIR –8 dB, T_{60} 0.6 s), the unprocessed score of 88.6% was reduced by one keyword (1.4% correct).

The use of an identical algorithm and training allowed an exact comparison to be made between traditional within-language conditions, in which the network is trained and tested both on the same language (Healy *et al.*, 2020), and the current cross-language examination (see Fig. 4). Objective measures showed that benefit was similar across these language conditions, with ESTOI and STOI benefit values both within a few percentage points. The linear mixed model failed to reveal a significant interaction between algorithm processing and language. Finally, the actual intelligibility benefit demonstrated by the current NH listeners hearing across-language conditions was highly comparable to that observed previously for within-language conditions, particularly following RAU transform.

RAU values are essentially equal to percent-correct values in the region free from floor and ceiling effects, from approximately 15% to 85% correct. The relationship between RAUs and percent correct diverges outside of this range, with 1% point corresponding to increasingly greater than 1 RAU. This expansion of RAU values counteracts the compression of percent-correct values against the floor or ceiling. Accordingly, they are a preferred comparison metric, particularly when floor or ceiling values are observed. This influence can be clearly observed currently. The data displayed in the left half of Fig. 4 are free of strong floor or ceiling effects, and the benefit difference across studies is approximately 2 points when expressed as raw percent correct or RAUs. In contrast, the data displayed in the right half of that figure contain a strong ceiling effect. This benefit difference across studies of 10 points (raw percent correct) becomes 1 point (RAU) when the compression of percent correct against the ceiling is addressed *via* the transform.

Also notable is that overall intelligibility of the target talker is higher in the current study relative to that of the within-language study of Healy *et al.* (2020). This is likely due to the fact that the Mandarin MSP test used a clear professional talker and contained sentences of fixed length containing familiar everyday monosyllabic words. In contrast, the target English-language materials employed previously (the Institute of Electrical and Electronics Engineers sentences, IEEE, 1969) are generally considered to be somewhat challenging. The IEEE sentences vary in length and contain words having relatively low frequency of occurrence. Further, they were produced by a typical nonprofessional talker, and so articulation was likely less toward clear speech. Benefit is correlated with unprocessed scores, which is expected because benefit values are derived in part from unprocessed scores. The current relationship ($|r| = 0.75$, $p < 0.0001$) was typical, and associated higher unprocessed scores with less benefit. Accordingly, higher baseline unprocessed scores worked against the current across-language benefit, and the observation of benefit comparable to that within a single language occurred despite this effect.

The current model is talker-independent and so employed different talkers for network training versus testing. Because these talkers were selected somewhat arbitrarily, and because the training versus test languages were selected to be maximally different, the current algorithm should be expected to generalize broadly across talkers and languages. To help confirm this, the current results were repeated using additional test talkers speaking another language. Four talkers (two female, two male) speaking Columbian Spanish from Guevara-Rukoz *et al.* (2020) were arbitrarily selected and formed into female-male pairs with the female arbitrarily assigned as the target in one pair and the male arbitrarily assigned as the target in the other pair. The mixing of equal-duration sentences at a TIR of –8 dB and the addition of room reverberation at $T_{60} = 0.6$ s was performed as described in Sec. II. Twenty sentence pairs were prepared for each talker pair, with all 80 sentences distinct from one another. The algorithm was trained using English-language materials as described in Sec. II. ESTOI averaged across talker pairs increased from 16.7% to 50.4% (unprocessed to processed), STOI increased from 37.3% to 67.2%, and PESQ increased from 0.77 to 1.73. The English-Spanish benefit (processed score minus unprocessed score) for each of these metrics was substantial but below those for English-Mandarin (ESTOI: English-Spanish benefit 33.7% points versus English-Mandarin benefit 41.2% points; STOI: 29.9 versus 31.8% points; PESQ: 0.96 versus 1.18; and ΔSDR: 12.4 versus 13.0 dB). It can therefore be concluded that the current cross-language generalization is not restricted to the selection of Mandarin as the generalization language or to the particular talkers who produced it.

Although the current study was performed using listeners with NH, listeners with typical sensorineural hearing loss represent a population of particular need for speaker separation/noise reduction/dereverberation. Prior deep learning algorithm studies from our laboratory that employed identical conditions for both hearing-impaired (HI) and NH listeners make it possible to compare benefit across these groups. Benefit observed for HI listeners was on average 4.5 times that observed for NH listeners, and in no study was that factor below 1.9 (Healy *et al.*, 2013; Healy *et al.*, 2017; Zhao *et al.*, 2018; Healy *et al.*, 2019; Healy *et al.*, 2020). So additional benefit is clearly anticipated in the current across-language conditions for HI listeners having typical hearing loss and who wear hearing aids.[3]

The two primary languages employed currently are among the most different known, with large etymological, orthographic, phonemic, and phonetic differences. However, because current performance was comparable to that observed within a single language, it must be considered that the learning accomplished by the neural network transcended these various aspects of language. It is therefore of potential interest to consider what commonalities exist across training and test that allowed the network to accomplish the current task.

At the root of all acoustic commonalities in speech is the human physiological speech-production mechanism. There are a limited number of ways that humans can manipulate their vocal folds and the resonances or constrictions of the oral and nasal cavities. Accordingly, there is a relatively limited array of buzzes, pops, and hisses that a human vocal tract can produce. So although the differences between languages with independent ancestry are vast, the acoustic signals themselves are apparently not sufficiently different to impede the deep neural network much at all. Further, the network was trained using English-language materials with no anticipation of the future need to generalize to an entirely different language. Despite this, the learning that took place appears to have been centered on attributes of human speech that transcend language.

V. CONCLUSIONS

The ability of a deep learning based speaker separation and dereverberation system to generalize to conditions vastly different from those experienced during network training was assessed. The current network was deep CASA, which employed Dense-UNet to separate talkers in each time frame, then a TCN to organize those frames over time (Fig. 1). Complex time-frequency masking allowed both the magnitude and phase of the target speech to be estimated. The challenging listening conditions involved two concurrent talkers and large amounts of room reverberation, and the network was tasked with isolating the target talker and removing reverberation (Fig. 2).

The generalizations required across training and test included: different utterances, TIRs, reverberation RIRs, speech corpora/recording channels, and talkers. Further, a perhaps ultimate generalization was introduced involving a different language, as training was conducted using English-language speech materials and testing was conducted using Mandarin-language materials. These languages are the world's two most commonly spoken and the lack of known common ancestry causes them to possess extensive linguistic differences.

Significant intelligibility increases resulting from algorithm processing were observed for the NH listeners in every condition, which averaged 43.5 raw percentage points and 48.2 RAU points (Fig. 3). This across-language algorithm benefit was highly comparable to within-language benefit observed in conditions that were otherwise identical (overall difference across studies of 6.2 raw percentage points or 1.7 RAU points, Fig. 4). Substantially greater benefit than that observed currently is expected for listeners with hearing loss, who are especially intolerant of interfering sounds and room reverberation.

The current results suggest that a well-designed and trained deep learning algorithm is capable of vast generalizations across highly different acoustic conditions. Given the talker-independent nature of the current model and the use of two very different languages, the current algorithm should, in theory, be able to increase intelligibility by separating any two voices in any language.

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., and R01 DC012048 to D.L.W., with manuscript preparation also supported by F32 DC019314 to E.M.J.). We gratefully acknowledge computing resources from the Ohio Supercomputer Center, algorithmic assistance from Yuzhou Liu, and subject-response scoring assistance from Yixuan Zhang and Ke Tan.

[1] We note that the opposite direction (train on Mandarin and test on English) was not assessed.

[2] For readers who prefer the more traditional analysis of variance (ANOVA), a three-way repeated-measures ANOVA was performed using the same dependent variable, fixed effects, and interactions. It showed a pattern of significance identical to the linear mixed-effects model, except that the two-way interaction between TIR and T_{60} was also significant, indicating that the effect of TIR on intelligibility depended in part on T_{60} [$F(1,9) = 8.1, p = 0.02, \eta^2_{\text{partial}} = 0.47$].

[3] A variety of techniques exist for selecting the voice of interest in a hearing aid or other device. It can be selected simply by using the more intense voice, which is the desired target in most communication settings, and which communication partners naturally work to remedy when they are not providing one another with the most intense signal. It can be selected with the assistance of directional microphones or more advanced techniques such as eye gaze or EEG-based attention monitoring. Alternatively, the mixture can be segregated and different sources assigned to different spatial locations or opposite ears.

Allen, J. B., and Berkley, D. A. (**1979**). "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943–950.

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

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

ANSI (**2010b**). *S12.60 (R2015), Acoustical Performance Criteria, Design Requirements, and Guidelines for Schools, Part 1: Permanent Schools* (Acoustical Society of America, New York).

Apoux, F., and Healy, E. W. (**2013**). "A glimpsing account of the role of temporal fine structure information in speech recognition," in *Basic Aspects of Hearing: Physiology and Perception*, edited by B. C. J. Moore, R. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel (Springer, New York).

Bai, S., Kolter, J. Z., and Koltun, V. (**2018**). "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv:1803.01271.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**, 1–48.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).

Chen, J., and Wang, D. L. (**2017**). "Long short-term memory for speaker generalization in supervised speech separation," J. Acoust. Soc. Am. **141**, 4705–4714.

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.

Cicchetti, D. V. (**1994**). "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," Psychol. Assess. **6**, 284–290.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (**2020**). *Ethnologue: Languages of the World*, 23rd ed. (SIL International, Dallas, TX).

Fu, Q.-J., Zhu, M., and Wang, X. (**2011**). "Development and validation of the Mandarin speech perception test," J. Acoust. Soc. Am. **129**, EL267–EL273.

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.

Guevara-Rukoz, A., Demirsahun, I., He, F., Shan-Hiu, C. C., Supheakmungkol, S., Pipatsrisawat, K., Gutkin, A., Butryna, A., and Kjartansson, O. (**2020**). "Crowdsourcing Latin American Spanish for low-resource text-to-speech," in *Proceedings of the 12th Conference on Language Resources and Evaluation*, May 13–15, Marseille, France, pp. 6504–6513.

Habets, E. (**2020**). ehabets/RIR-Generator: RIR Generator (v2.2.20201022). Zenodo. https://doi.org/10.5281/zenodo.4117640 (last viewed 28 September 2021).

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., 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. (**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., and Warren, R. M. (**2003**). "The role of contrasting temporal amplitude patterns in the perception of speech," J. Acoust. Soc. Am. **113**, 1676–1688.

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.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (**2016**). "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proceedings of ICASSP*, May 20–25, Shanghai, China, pp. 31–35.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (**2017**). "Densely connected convolutional networks," in *Proceedings of CVPR*, July 21–26, Honolulu, HI, pp. 2261–2269.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

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. Process. **24**, 2009–2022.

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.

Kolbaek, M., Yu, D., Tan, Z. H., and Jensen, J. (**2017**). "Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," IEEE/ACM Trans. Audio Speech Lang. Process. **25**, 1901–1913.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2020**). "lmerTest: Tests in Linear Mixed Effects Models, R package version 3.1-3," https://CRAN.R-project.org/package=lmerTest (last viewed 28 September 2021).

Lea, C., Vidal, R., Reiter, A., and Hager, G. D. (**2016**). "Temporal convolutional networks: A unified approach to action segmentation," in *Proceedings of ECCV*, October 11–14, Amsterdam, the Netherlands, pp. 47–54.

Liu, Y., and Wang, D. L. (**2019**). "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," IEEE/ACM Trans. Audio Speech Lang. Process. **27**, 2092–2102.

McGraw, K. O., and Wong, S. P. (**1996**). "Forming inferences about some intraclass correlation coefficient," Psych. Methods **1**, 30–46.

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.

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

Paul, D. B., and Baker, J. M. (**1992**). "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the Workshop on Speech and Natural Language*, February 23–26, Harriman, New York, pp. 357–362.

R Core Team (**2020**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria; https://www.R-project.org/ (last viewed 30 March 2021).

Ranbom, L. J., and Connine, C. M. (**2007**). "Lexical representation of phonological variation in spoken word recognition," J. Mem. Lang. **57**, 273–298.

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*, May 7–11, Salt Lake City, UT, pp. 749–752.

Ronneberger, O., Fischer, P., and Brox, T. (**2015**). "U-Net: Convolutional networks for biomedical image segmentation," arXiv:1505.04597.

Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. Lond. B. **336**, 367–373.

Satterthwaite, F. E. (**1941**). "Synthesis of variance," Psychometrika **6**, 309–316.

Souza, P. (**2016**). "Speech perception and hearing aids," in *Hearing Aids*, edited by G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper (Springer, New York).

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.

Underhill, A. (**2005**). *Sound Foundations*, 2nd ed. (Macmillan, New York).

Vincent, E., Gribonval, R., and Févotte, C. (**2006**). "Performance measurement in blind audio source separation," IEEE Trans. Audio Speech Lang. Process. **14**, 1462–1469.

Wang, D. L., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (Wiley-IEEE, Hoboken, NJ).

Wang, D., and Zhang, X. (**2015**). "THCHS-30: A free Chinese speech corpus," arXiv:1512.01882.

Williamson, D. S., Wang, Y., and Wang, D. L. (**2016**). "Complex ratio masking for monaural speech separation," IEEE/ACM Trans. Audio Speech Lang. Process. **24**, 483–492.

World Health Organization. (**2020**). "Deafness and hearing loss, Fact Sheet," https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (last viewed 28 September 2021).

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.