Loughborough University

This item was submitted to Loughborough's Research Repository by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Predicting speech perception in older listeners with sensorineural hearing loss using automatic speech recognition

PLEASE CITE THE PUBLISHED VERSION

https://doi.org/10.1177/2331216520914769

PUBLISHER

SAGE Publications

VERSION

VoR (Version of Record)

PUBLISHER STATEMENT

This is an Open Access Article. It is published by SAGE under the Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC 4.0). Full details of this licence are available at: https://creativecommons.org/licenses/by-nc/4.0/

LICENCE

CC BY-NC 4.0

REPOSITORY RECORD

Fontan, Lionel, Tom Cretin-Maitenaz, and Christian Fullgrabe. 2020. “Predicting Speech Perception in Older Listeners with Sensorineural Hearing Loss Using Automatic Speech Recognition”. Loughborough University. https://hdl.handle.net/2134/13618916.v1.

Check for updates

Original Article

Predicting Speech Perception in Older Listeners with Sensorineural Hearing Loss Using Automatic Speech Recognition

Trends in Hearing
Volume 24: 1–16
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2331216520914769
journals.sagepub.com/home/tia
SAGE

Lionel Fontan[1], Tom Cretin-Maitenaz[2,3], and Christian Füllgrabe[4]

Abstract

The objective of this study was to provide proof of concept that the speech intelligibility in quiet of unaided older hearing-impaired (OHI) listeners can be predicted by automatic speech recognition (ASR). Twenty-four OHI listeners completed three speech-identification tasks using speech materials of varying linguistic complexity and predictability (i.e., logatoms, words, and sentences). An ASR system was first trained on different speech materials and then used to recognize the same speech stimuli presented to the listeners but processed to mimic some of the perceptual consequences of age-related hearing loss experienced by each of the listeners: the elevation of hearing thresholds (by linear filtering), the loss of frequency selectivity (by spectrally smearing), and loudness recruitment (by raising the amplitude envelope to a power). Independently of the size of the lexicon used in the ASR system, strong to very strong correlations were observed between human and machine intelligibility scores. However, large root-mean-square errors (RMSEs) were observed for all conditions. The simulation of frequency selectivity loss had a negative impact on the strength of the correlation and the RMSE. Highest correlations and smallest RMSEs were found for logatoms, suggesting that the prediction system reflects mostly the functioning of the peripheral part of the auditory system. In the case of sentences, the prediction of human intelligibility was significantly improved by taking into account cognitive performance. This study demonstrates for the first time that ASR, even when trained on intact independent speech material, can be used to estimate trends in speech intelligibility of OHI listeners.

Keywords

automatic speech recognition, speech intelligibility, age-related hearing loss, suprathreshold auditory processing, cognition

Received 21 June 2019; revised 16 February 2020; accepted 2 March 2020

To measure the impact of age-related hearing loss (ARHL) on communicative function, as well as to quantify improvements in speech processing following auditory rehabilitation through hearing aids (HAs), tests of speech intelligibility—clinically referred to as "speech audiometry"—have long been used as a complement to pure-tone audiometry (Fournier, 1951; Hirsh et al., 1952; Hudgins et al., 1947). Generally, either the proportion of correctly identified utterances at one or several constant presentation or speech-to-noise level(s) is determined, or the so-called speech reception threshold (SRT; American National Standard Institute [ANSI], 1969), corresponding to the presentation or speech-to-noise level required to achieve a given performance level (e.g., 50% correct), is adaptively tracked. However, speech audiometry is less reliable and more time-consuming than would be desirable for clinical practice. For example, in France, HA audiologists often evaluate speech intelligibility in a given listening situation using 10-word lists such as those developed by Fournier (1951). According to Moulin et al. (2016), at

[1]Archean LABS, Montauban, France
[2]Service d'Oto-Rhino-Laryngologie, d'Oto-Neurologie et d'ORL Pédiatrique, Centre Hospitalier Universitaire de Toulouse, France
[3]Ecole d'Audioprothèse de Cahors, Université Paul Sabatier Toulouse III, France
[4]School of Sport, Exercise and Health Sciences, Loughborough University, UK

Corresponding Author:
Lionel Fontan, Archean LABS, 20 Place Prax-Paris, Montauban, France.
Email: lfontan@archean.tech

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

least five of such lists should be administered for each test condition to achieve a reliable estimate of speech intelligibility. In the context of fitting HAs and their fine-tuning, where several settings for different processing features (e.g., gain rule, dynamic range compression, noise reduction) need to be explored, the repetition of the test procedure for each combination of settings could therefore become tedious for the listener. Especially for older patients, this might lead to an increase in fatigue and a decrease in attention, and, subsequently, yield lower levels of and higher variability in performance over the course of the assessment. In addition, because familiarity with the speech material affects test performance (e.g., Hustad & Cahill, 2003), the test material needs to be refreshed for each test condition, and the limited number of test items restricts the number of test conditions that can be assessed within the same patient without repeating the test material.

The use of speech-intelligibility prediction systems could overcome these shortcomings. For example, in the fields of telecommunications and room acoustics, a number of predictive models of speech intelligibility have been developed (e.g., Speech Intelligibility Index, ANSI, 1997; Articulation Index, French & Steinberg, 1947; Speech Transmission Index, Steeneken & Houtgast, 1980). In these models, the most important acoustic features of the communication channels that influence speech intelligibility are assessed (such as signal-to-noise ratio [SNR] and reverberation time) and used to predict intelligibility scores that listeners would obtain under the same acoustic conditions. As a consequence, the reliability of these models for predicting intelligibility performance under acoustic conditions other than those used to collect human reference data (e.g., with other types of noise or different SNRs) is uncertain (for a more detailed discussion of these models, see Schädler et al., 2015; for a review of current speech-intelligibility and speech-quality prediction models, see Falk et al., 2015).

In contrast, automatic speech recognition (ASR) systems could constitute a more (yet not entirely) reference-free means for predicting speech intelligibility. Indeed, ASR has been used successfully to estimate the intelligibility of speech degraded by the presence of background noise (Barker & Cooke, 2007; Spille et al., 2018) or speech pathologies (e.g., Fontan, Pellegrini, et al., 2015; Maier et al., 2009).

By extension, and in case that the perceptual consequences of ARHL can be accurately simulated by signal processing, ASR should also be usable for the prediction of speech intelligibility for listeners with ARHL. From a practical perspective, this would mean that HA audiologists could calculate rapidly and at no "cost" to the patient the ASR-predicted intelligibility performance for any number of combinations of listening conditions (e.g., in quiet or in different types of noises at different levels, with or without HAs, and for various combinations of HA settings).

Recent work by Kollmeier and colleagues has shown encouraging results in this regard (Kollmeier et al., 2016; Schädler et al., 2015). Kollmeier et al. (2016) used an ASR system to predict SRTs for matrix sentences in noise for a large group of hearing-impaired (HI) listeners whose ages ranged between 23 and 82 years. They modeled the effect of ARHL on speech intelligibility using Plomp (1978)'s framework, in which any hearing loss is considered as the combination of a "hearing loss of class A" (where A stands for attenuation) and a "hearing loss of class D" (where D stands for distortion). The former results in an upward shift of absolute thresholds, and, thus, can be compensated for by increasing the presentation level of the speech signal, while the latter represents the temporal and spectral distortions that affect the intelligibility of speech independently of its audibility. Such distortions include the loss of frequency selectivity (FS; reduction in the ability to resolve spectral components; e.g., Baer & Moore, 1993) and loudness recruitment (LR; reduction of the intensity dynamic causing an exaggerated perception of intensity changes and an intolerance to loud noises; e.g., Moore, 2007). The attenuation component (A) was simulated by using a thresholding procedure in the ASR system based on each listener's audiogram. The distortion component (D) was estimated based on the mismatch between the listeners' intelligibility scores and those predicted by the ASR system when only A was taken into account. The effect of D on intelligibility was then simulated by an additive white Gaussian noise that was added to the original speech signal before feeding the speech-and-noise mixture once again to the ASR system. Correlation coefficients between human and machine SRTs ranged from 0.56 to 0.84 (for a stationary noise) and from 0.69 to 0.91 (for a fluctuating noise). In comparison, predictions based on the Speech Intelligibility Index (ANSI, 1997) yielded correlations coefficients that did not exceed 0.77 and 0.72 for the stationary and fluctuating noise, respectively. However, the approach taken in that study was far from being reference-free, as the ASR system was trained and tested on the same speech materials (i.e., the German matrix sentences at different SNRs).

A different approach was taken by Fontan et al. (2014, 2017). First, they used different speech materials for training and testing the ASR system. Second, they simulated not only elevated hearing thresholds (HTs) but also some suprathreshold auditory processing deficits occurring in listeners with ARHL. They used an algorithm developed by Nejime and Moore (1997) that takes audiometric thresholds as the input to also simulate the effects of FS loss and LR. For example, in

Fontan et al. (2017), different speech materials were processed through this ARHL simulator, using nine audiograms that are typical for ages ranging from 60 to 110 years and represent increasing levels of severity of ARHL (Cruickshanks et al., 1998). The processed speech was then presented to an ASR system and 60 young normal-hearing participants. The results revealed very strong positive correlations between human and machine identification scores (all $r \geq .90$), indicating that trends in human intelligibility as a function of the simulated degrees of hearing loss could be accurately predicted by ASR.

However, Fontan et al. (2017) compared machine scores to the *average* intelligibility scores obtained by the 60 participants in each of the 9 ARHL-simulation conditions. It is therefore not clear if ASR can also be used to predict intelligibility for *individual* cases of ARHL, which are likely to show a greater variability in audiometric profile, suprathreshold auditory processing, and cognitive functioning.

Also, the application of this prediction method to older listeners with actual (i.e., not simulated) ARHL assumes that the signal processing used to simulate the perceptual consequences of ARHL is accurate and that the list of simulated auditory processing deficits occurring in ARHL is exhaustive. However, there is increasing evidence that older listeners also present altered sensitivity to temporal-envelope (e.g., Füllgrabe et al., 2003, 2015; He et al., 2008) and temporal-fine-structure information (e.g., Füllgrabe, 2013; Füllgrabe et al., 2017; Grose & Mamo, 2010; Ross et al., 2007; for a review, see Füllgrabe & Moore, 2018) and that those deficits are associated with poorer speech perception (e.g., Füllgrabe et al., 2015; Neher et al., 2011). Such age-related temporal suprathreshold processing deficits are not simulated by the algorithm of Nejime and Moore (1997). In addition, individual variations and age-related changes in some (but not all; Füllgrabe & Rosen, 2016b) linguistic and cognitive abilities are associated with speech-perception performance (e.g., Carroll et al., 2016; Füllgrabe et al., 2015). As with increasing age more people are subject to cognitive decline (e.g., Baltes and Lindenberger, 1997; Verhaeghen & Salthouse, 1997), a number of older listeners might find themselves with insufficient cognitive resources to perform optimally complex tasks such as speech perception (Füllgrabe & Rosen, 2016a). This age-related effect is not taken into account by the ASR system.

Fontan et al. (2017) simulated the combined effects of three components of ARHL, namely (a) the elevation of HTs, (b) the loss of FS, and (c) LR. Thus, the relative effect of each of these components on the precision of the prediction of speech intelligibility is currently not known.

The present study addresses the issues outlined earlier by comparing intelligibility scores obtained by individual older listeners with ARHL with the predictions of an ASR system fed with speech signals processed to simulate *for each individual listener* one (HT elevation alone), two (HT elevation + FS loss; HT elevation + LR), or all three (HT elevation + FS loss + LR) of the perceptual consequences of ARHL.

While the long-term objective is to use ASR to predict aided speech intelligibility in a range of listening environments (e.g., in quiet, in the presence of different background sounds), it was decided to limit the scope of the present study to the prediction of *unaided* perception performance for speech presented *in quiet*. This was done because the listener-related variability in suprathreshold auditory and cognitive processing abilities are likely to play a larger role when speech has to be understood in the presence of background noise and/or after being processed through a HA (e.g., Stone et al., 2009). In addition, in France, speech audiometry is often performed in quiet, as shown by a recent survey of the current audiological practice of French HA audiologists (Rembaud et al., 2017).

Methodology

Participants

Recruitment. Twenty-eight older ($\geq$60 years) native French speakers, presenting for their first consultation at the ENT department of the Honoré Cave Hospital or a local HA dispensing center (Montauban, France), were recruited, based on their having relatively symmetrical and sloping high-frequency hearing losses, as is typical for mild-to-moderate ARHL. The difference in average hearing sensitivity in the low-frequency ($\leq$1 kHz) and high-frequency ($\geq$4 kHz) regions was at least 20 dB. Extreme cases of (nearly) normal hearing sensitivity or of more severe hearing losses (for the latter, the presence of cochlear dead regions becomes more likely) were not considered for participation to reduce the possibility of observing ceiling and floor effects on the speech-identification tests. None of the participants self-reported having been excessively exposed to loud sounds during their life. While some of the participants self-reported experiencing tinnitus, which can affect speech intelligibility (Ryu et al., 2012), none of them judged its presence as having a deleterious effect on their ability to understand speech. Only people with no prior experience with speech-intelligibility tests were invited to participate to ensure that familiarity with the speech materials used in the present study did not affect performance. All participants were confirmed as right-handed based on their results on the Edinburgh Handedness Inventory (Oldfield, 1971) and had normal

or corrected-to-normal vision. Prior to the start of the study, approved by the ethical committee of the Honoré Cave Hospital (Montauban, France), all participants provided informed written consent.

Cognitive Screening. To minimize the possibility that test performance might be affected by pathological cognitive status, the French version of the Mini-Mental State Examination (MMSE; Kalafat et al., 2003), a frequently used screening tool for the assessment of cognitive impairment, was administered to all participants. Two of the recruited participants scored less than 27 points (out of the maximum of 30), which has been considered as outside the range of normal cognitive functioning (e.g., Bassuk et al., 2000; Bruce et al., 1995; Zaudig, 1992), and hence were excluded from the study. As expected based on population norms (Crum et al., 1993), MMSE scores declined with increasing age (Spearman's $rho = -.52$, $p = .005$, one-tailed).

Audiometric Assessment. For the remaining 26 participants, pure-tone audiometry was reconducted in the test (i.e., right) ear, following the guidelines of the British Society of Audiology (2011). An Interacoustics Affinity 2.0 AC440 audiometry module with a 3M Peltor H7A headset and a RadioEar B-81 bone transducer were used for air- and bone-conduction audiometry, respectively. All testing was carried out in a sound-treated room complying with the requirements set out by the French public health code for audiometric booths (République Française, 2017); the measured ambient noise level over a 1-hr period was 32 dB(A) compared with the maximum acceptable level of 40 dB(A).

Air-conduction audiometric thresholds were measured at octave frequencies between 0.125 and 8 kHz, as well as at 0.75, 1.5, 3, and 6 kHz. To increase the precision of the threshold estimate, and therefore the precision of the ASR-based prediction based on the simulation of the participants' hearing losses, the final step size for the adaptive procedure was set to 2 dB instead of the routinely used 5 dB.

Bone-conduction audiometric thresholds were measured at octave frequencies between 0.25 and 4 kHz, as well as 0.75, 1.5, and 3 kHz. As recommended by the British Society of Audiology (2011), a masking (white) noise was systematically presented to the nontest (i.e., left) ear at 15 dB above the presentation level of the pure tones to minimize the contribution of this ear to the bone-conduction thresholds. The sensorineural nature of the participants' hearing losses was confirmed by air-bone gaps $\leq$12 dB at each test frequency. Two participants, who had larger air-bone gaps, were excluded from the study.

The individual and mean audiograms of the remaining 24 participants (9 females; mean age = 71.3 years, standard deviation [SD] = 7.9), who fulfilled all inclusion criteria, are shown in Figure 1. Mean audiometric thresholds declined progressively with increasing frequencies. The mean audiogram spanned a very similar range of hearing sensitivities (i.e., from 18 dB hearing level [HL] at 0.125 kHz to 66 dB HL at 8 kHz) to that extrapolated for an age of 72 years by Fontan et al. (2017; see their Figure 1) on the basis of the audiometric data reported by Cruickshanks et al. (1998), averaged across gender and ears. However, the overall sensitivity averaged across all frequencies of 42 dB HL was 9-dB higher than that reported by Fontan et al. (2017). The pure-tone average (PTA) for frequencies of 0.5, 1, 2, and 4 kHz did not significantly correlate with the age of the participants ($r = .09$, $p = .342$, one-tailed). Individual characteristics in terms of demographic, audiometric, and cognitive data for the 24 participants are shown in Table 1.

Human Speech Intelligibility

Speech Materials. Human speech intelligibility was assessed in quiet for the three types of speech materials most frequently used for speech audiometry by HA audiologists in France (i.e., logatoms, words, and sentences; Rembaud et al., 2017). All stimuli were taken from the recordings produced by the Collège National d'Audioprothèse (CNA; 2007), a French nonprofit organization of HA audiologists providing technical, pedagogic, and deontological guidance and recommendations, as well as promoting scientific research in the domain of audiology (http://www.college-nat-audio.fr). All stimuli

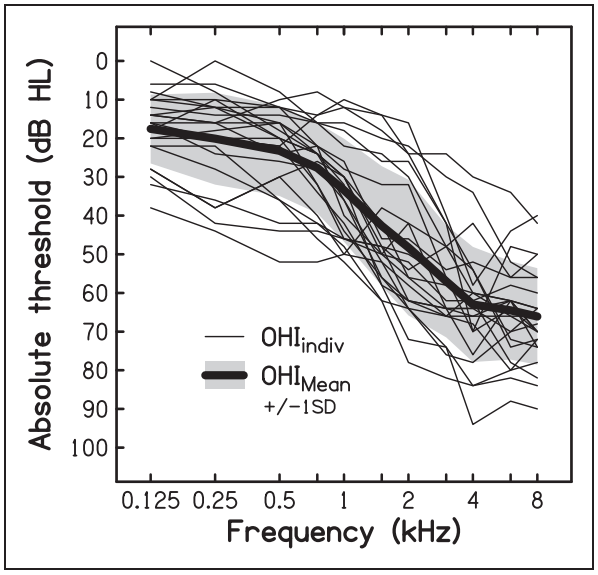


Figure 1. Results of Pure-Tone Air-Conduction Audiometry for the Test (i.e., Right) Ears of the 24 OHI Participants. The thin lines represent the individual audiograms. The thick line and associated gray-shaded area represent the mean audiogram ±1 SD.
OHI = older hearing-impaired; SD = standard deviation.

Table 1. Individual Characteristics for the 24 Older Hearing-Impaired (OHI) Participants in Terms of Gender (F = Female, M = Male), Age (Years), MMSE Score (Out of 30), and Pure-Tone Average for Audiometric Frequencies of 0.5, 1, 2, and 4 kHz (PTA; in dB HL) for the Test (i.e., Right) Ear.

Participant	Gender	Age	MMSE	PTA
OHI_1	F	73	29	19.5
OHI_2	F	69	27	21
OHI_3	H	68	28	24
OHI_4	H	71	29	27.5
OHI_5	H	67	29	28.5
OHI_6	H	73	29	32
OHI_7	F	75	29	36
OHI_8	F	60	30	36.5
OHI_9	F	87	30	37
OHI_{10}	H	62	29	40.5
OHI_{11}	H	61	30	41.5
OHI_{12}	H	61	29	45.5
OHI_{13}	H	66	28	45.5
OHI_{14}	H	85	27	45.5
OHI_{15}	H	70	29	46
OHI_{16}	F	89	27	48
OHI_{17}	F	67	29	50
OHI_{18}	H	73	28	51
OHI_{19}	H	77	27	52
OHI_{20}	F	77	28	53
OHI_{21}	F	78	27	54
OHI_{22}	H	67	30	54.5
OHI_{23}	H	67	29	57
OHI_{24}	H	68	29	61.5
OHI_{Mean}	(9F/15M)	71.3	28.6	42.0

Note. Participants are ranked in increasing order of PTA. MMSE = Mini-Mental State Examination; PTA = pure-tone average.

were pronounced by the same adult male native speaker of standard French and recorded using a 44.1-kHz sampling rate and 32-bit quantization.

For the identification of logatoms, the nonsense vowel–consonant–vowel (VCV) stimuli from Lists 1 to 4 of the test material, provided by the CNA (2007) and originally developed by Dodelé and Dodelé (2000), were used. Each of these VCV stimuli was composed of 1 of the 17 most frequent French consonants (C =/p,t,k,b,d, g,m,n,f,s,ʃ,v,z,ʒ,ʁ,l,w/), presented within a given vocalic context composed of two different vowels (V =/i,e,ɛ,ɛ̃,ø, o,ɔ,ɔ̃,a,ɑ̃,y,u/; e.g., "iza," "ato"). Each list contained all consonants once (i.e., 17 logatoms), for a total of 68 logatoms used in the study.

For the identification of words, 60 disyllabic masculine nouns (corresponding to 6 of the 40 ten-word lists developed by Fournier, 1951), each preceded by the French masculine definite article "le", were used (e.g., "le soldat"—"the soldier"). Although these lists are not phonetically balanced, they are exclusively composed of nouns starting with a consonant and ending with any vowel other than /ə/, which has a special phonological status in the French language, to minimize differences in phonological structure between items within and across lists (Fournier, 1951).

For the identification of sentences, 40 sentences from the French version of the Hearing in Noise Test (HINT; Vaillancourt et al., 2005) were used. The sentences are rather simple and somewhat predictable, with each sentence being composed of a single assertive clause combining words chosen from a language comprehension test for children aged 6 to 7 years (Leduc, 1997; e.g., "Le camion est rouge."—"The truck is red.").

General Procedure. All participants were tested with the three types of speech materials, with the order of the identification tests being counterbalanced across participants. Prior to data collection with each test, participants were briefly familiarized with the test stimuli and procedure using additional practice stimuli that were not used during the test phase: one list of logatoms, one list of words, and four sentences.

Stimuli were presented at 50 dB sound pressure level (SPL) through an Interacoustics Affinity 2.0 audiometer connected to a 3M Peltor H7A headset to the participant seated in the same sound-treated room as that used for the audiometric assessment. This level of presentation (corresponding to "quiet speech" in the Affinity graphical user interface) was chosen, as informal preliminary testing of other older patients with a similar range of ARHLs revealed a wide range of speech-identification performance while avoiding large floor and ceiling effects. It is noteworthy that this level is lower by approximately 15 dB than that associated with a normal conversational level. The rationale for assessing speech-identification performance at a fixed presentation level was that ASR systems, in contrast to human listeners, use signal intensity normalization techniques for processing and modeling speech. In case of an adaptive procedure tracking the SRT, this would differently affect human and machine intelligibility scores for speech presented in quiet. Finally, as both ASR and the ARHL-simulation program process mono signals, it was decided to present the speech stimuli monaurally to the right ear.

Participants were instructed to report back verbally the VCVs or words they had heard and to guess in case they were uncertain. No feedback on the test performance was provided, but all participants received verbal encouragements from the experimenter to maintain motivation and to reduce possible frustration in those participants struggling with the task. Responses were recorded using a microphone positioned in front of the participant for off-line manual transcription based on the International Phonetic Alphabet and scoring. For each participant, the final intelligibility score corresponded to the percentage of entirely correctly

identified VCVs (for the logatom-identification test) or words (for the word- and sentence-identification tests).

Machine Speech Intelligibility

ASR System. In the present study, the SPHINX-3 engine (Seymore et al., 1998), an open-source speech recognizer based on hidden Markov models (Rabiner, 1989), was used. Acoustic models designed to process 16-kHz speech recordings and representing 35 phones and five kinds of pauses were created (i.e., trained) using a 31-hr-long corpus of French radio broadcast recordings (Galliano et al., 2009), containing utterances from approximately 600 speakers. The acoustic features were extracted using a 16-ms sliding window. Each feature was composed of 12 mel-frequency cepstral coefficients (MFCCs; Davis & Mermelstein, 1980), calculated in the 0.3-to-8-kHz range, and the signal energy, as well as the first and second derivatives of these 13 values (12 MFCCs + the signal energy), resulting in a total of 39 dimensions. Each phone was modeled by a hidden Markov model representing the probability of transitions between three acoustic states (such states include the attack, sustain, and release parts of phones), and the distribution of acoustic features for each state was represented by a Gaussian mixture model with 32 Gaussian models (Deléglise et al., 2005; Estève, 2009).

As the size of the lexicon used in the ASR system influences the probabilities of occurrence of the target items, which might affect the correlations with human intelligibility scores, three lexicons varying in size were created for each of the three speech materials used in this study (referred to in the remainder of the article as "small", "medium", and "large" lexicons). The small lexicon was only comprised of the stimuli that were presented to the participants during the identification tasks. Given its high specificity, compared with the two other lexicons, the small lexicon was expected to yield the highest machine intelligibility.

For the logatom-identification test, one finite-state grammar was created, containing only one final state (i.e., a single "word" to be recognized) that could be actualized under the form of any of (a) the 68 test logatoms presented to the participants (small lexicon), (b) the 68 test logatoms and the 17 logatoms that were used for the training of the participants (medium lexicon), and (c) the 2,448 VCVs that can be obtained by combining the 17 consonants and 12 vowels used in the logatom-identification test (large lexicon).

For the word-intelligibility test, a bigram language model was used. This model contained sequences of two words beginning by the French masculine definite article "le" followed by a masculine noun. The probability associated with each noun corresponded to the probability of their occurrence in spoken French, as described by New et al. (2007). The bigram model was associated with a lexicon consisting of (a) the 60 disyllabic nouns from the test of Fournier (1951) that were presented to the participants (small lexicon), (b) 300 words from the 30 lists of Fournier (1951)'s dissyllabic words distributed by the CNA (2007; medium lexicon), or (c) the 6,491 masculine nouns starting with a consonant included in the lexicon created by de Calmès et al. (2005; large lexicon).

For the sentence-identification test, a trigram model was created based on the ESTER2 corpus (Deléglise et al., 2005; Galliano et al., 2009). The trigram model was associated with a lexicon comprising (a) the 129 words included in the 40 HINT sentences that were presented to the participants (small lexicon), (b) the 566 words contained in the 100 HINT sentence recordings provided by the CNA (2007; medium lexicon), or (c) the whole lexicon created by de Calmès et al. (2005), containing 62,351 French words (large lexicon).

To check that the ASR system achieved a sufficiently high performance (i.e., more than 80% correct identification) with "normal" (i.e., undegraded) speech, machine intelligibility scores for the three unprocessed speech materials were computed, using the small lexicons. Performance for logatoms, words, and sentences was 84.7, 98.3, and 90.8%, respectively, indicating that the ASR system was working as expected.

Simulation of Hearing Loss. The algorithm described by Nejime and Moore (1997) was used to simulate some of the perceptual consequences of ARHL. The algorithm was implemented in a custom-written MATLAB program. Based on the audiometric thresholds provided to the program, three effects associated with ARHL were simulated: (a) elevated HTs (by attenuating the frequency components in several frequency bands according to the threshold values given as an input); (b) reduced FS (by spectrally smearing the speech signal; Baer & Moore, 1993); and (c) LR (by raising the signal envelope to a power; Moore & Glasberg, 1993).

To simulate the elevation of HTs, the program uses frequency-dependent linear attenuation filters. The gain value for each filter is defined according to the corresponding audiometric threshold.

To simulate the loss of FS, the program first defines the degree of hearing loss based on the PTA for audiometric frequencies between 2 and 8 kHz, using three categories: "mild" ($15\,\text{dB HL} \leq \text{PTA}_{2\text{–}8\,\text{kHz}} < 35\,\text{dB HL}$), "moderate" ($35\,\text{dB HL} \leq \text{PTA}_{2\text{–}8\,\text{kHz}} < 56\,\text{dB HL}$), and "severe" ($\text{PTA}_{2\text{–}8\,\text{kHz}} \geq 56\,\text{dB HL}$). Depending on the category, a different degree of spectral smearing is applied to the power spectrum, using the algorithm described by Baer and Moore (1993). As auditory filters tend to broaden asymmetrically as a function of the degree of ARHL (e.g., Glasberg & Moore, 1986; Tyler et al., 1984), with more broadening generally occurring on the lower slope of the filters, higher broadening

factors were set for the lower (L) slopes of the simulated auditory filters than for the upper (U) slopes: Broadening factors were (L = 1.6, U = 1.2), (L = 2.4, U = 1.6), and (L = 4.0, U = 2.0) for the filters associated with a mild, moderate, and severe loss, respectively.

Finally, the program simulates LR by raising the normalized envelope of the speech signals (Moore & Glasberg, 1993). Depending on the degree of hearing loss to be simulated, the envelope is raised with more or less power.

To investigate the relative importance of each of the components of the ARHL-simulation algorithm for the prediction of human speech-identification performance, four simulation conditions were set in this study: HT elevation, HT elevation + FS loss, HT elevation + LR, and HT elevation + FS loss + LR.

Machine Intelligibility Score. Every item (VCV or word) was considered as correct if it was chosen as the most probable item by the ASR system. Final scores for each test corresponded to the percentage of correct items (in the case of sentences, every word of each sentence was taken into account).

Results

Human Speech Intelligibility

Intelligibility scores for each of the three speech materials are shown in Figure 2. In all three tests conducted at the same presentation level of 50 dB SPL, individual scores varied widely across participants, spanning almost the entire possible performance range. The distributions of performance for logatoms and sentences looked somewhat bimodal with identification scores falling either above 80 or below 50% correct. Twenty-six, 13, and 13% of the participants were unable to perform the logatom, word, and sentence identification tests, respectively, while 17% performed at ceiling in the sentence identification test. A Kolmogorov–Smirnov test showed that intelligibility scores for logatoms and sentences were not normally distributed (both $p < .05$). Hence, nonparametric tests were used in all subsequent inferential statistical analyses involving intelligibility scores. As expected, median performance (shown by the thick black lines) differed across the three speech materials: It was lowest for the meaningless logatoms (25.7%) and increased for the words in isolation (73.3%) to reach a near-ceiling level for the words in sentences (85.8%).

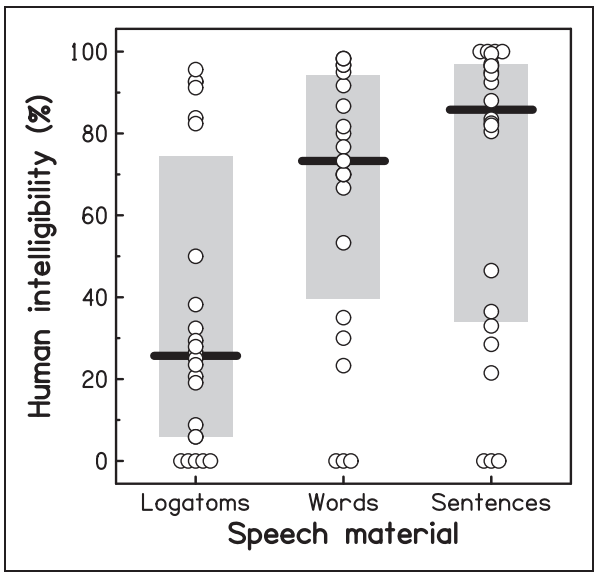


Figure 2. Speech Intelligibility for Logatoms, Words, and Sentences. Circles represent individual scores for the 24 OHI participants. Overlapping data points at the extremes are displaced horizontally for better visibility. The thick black line indicates median performance, and the gray area represents the associated interquartile range.

Prediction of Human Speech Intelligibility

Intelligibility scores for all three speech materials declined with age (Spearman's *rho* ranged from –.20 to –.28), but this trend was not significant (all $p \geq .091$, one-tailed). On the other hand, PTA correlated strongly with speech intelligibility, with higher PTAs being associated with lower intelligibility scores (all Spearman's *rho* $\leq$–.87, all $p < .001$, one-tailed). In the upper row of Figure 3, human intelligibility is plotted as a function of PTA for each of the three speech materials (see different panels). A nonlinear regression, using a generalized logistic function, was used to model the psychometric curves for each speech material. Consistent with previous observations (e.g., Kryter, 1994; Pichora-Fuller, 2008; Sheldon et al., 2008), the slope of the psychometric function becomes steeper, and the function shifts toward higher PTAs as the speech material contains more linguistic information.

The use of lexicons of different sizes resulted in changes in performance of the ASR system but hardly affected its ability to predict the trends in human speech intelligibility (see Table 2). For all three speech materials, highest machine intelligibility was achieved for the smallest lexicon. For logatoms, using the smallest lexicon yielded median intelligibility (30.6%) that was close to that observed in the older hearing-impaired (OHI) participants (25.7%). Machine intelligibility remained approximately the same when the used speech material consisted of isolated words (36.7%) or sentences (32.2%), thereby dramatically underestimating human intelligibility (73.3 and 85.8% for words and sentences, respectively). There is however a noticeable exception to this general trend: The predicted intelligibility scores for one OHI listener were always much higher than the

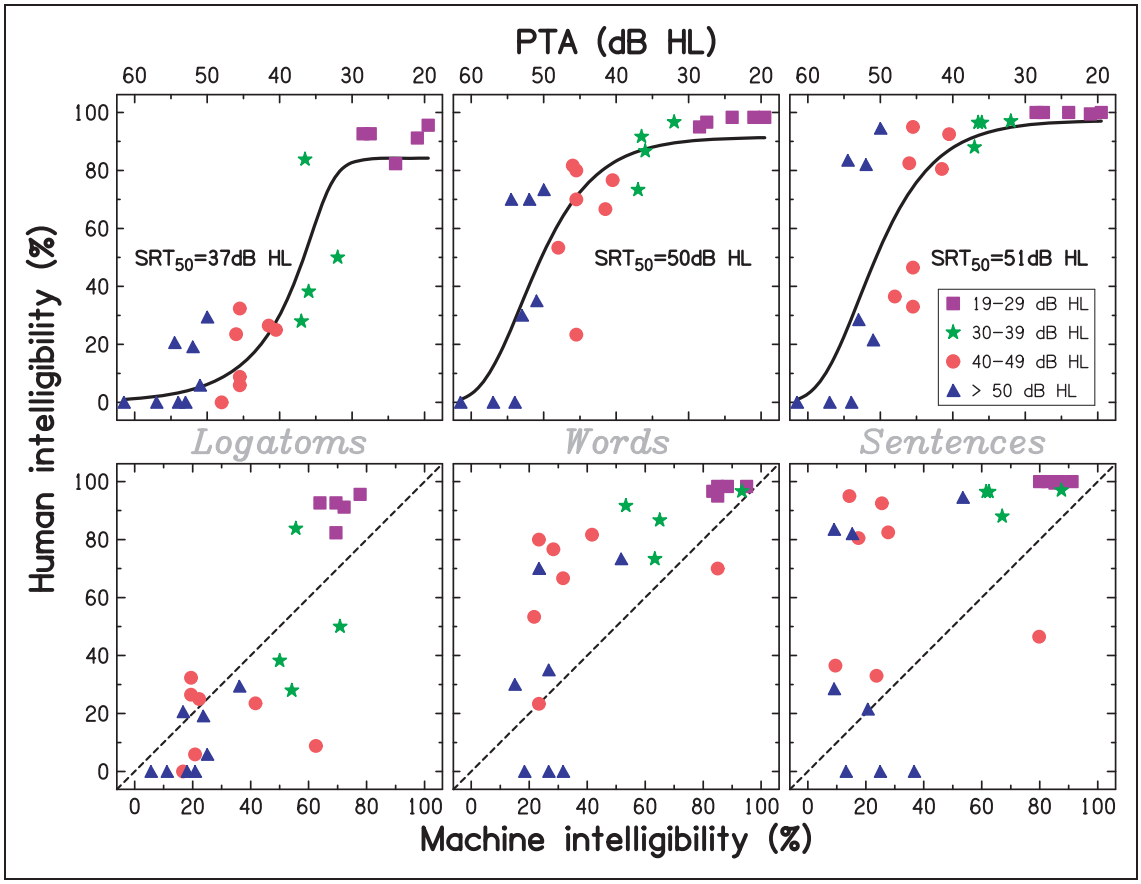


Figure 3. Human Intelligibility Scores Plotted as a Function of PTA and Machine Intelligibility Scores. Upper row shows scatterplots relating human intelligibility scores to the PTA computed over frequencies 0.5, 1, 2, and 4 kHz for each participant, and lower row shows scatterplots relating human intelligibility scores to individual scores predicted by the ASR system with a small lexicon and using the ARHL simulation implementing elevation of hearing thresholds, loss of frequency selectivity, and loudness recruitment. Each panel shows the results for a given speech material: logatoms (left panel), words (middle panel), and sentences (right panel). The different colors indicate different ranges of PTAs (see symbol legend).
PTA = pure-tone average; SRT = speech reception threshold.

Table 2. Median Machine Intelligibility (in %) and Correlation Between Human and Machine Intelligibility Scores (Spearman's *Rho*, One-Tailed), Using the Small, Medium, and Large Lexicon (Rows) for Each of the Three Speech Materials (Columns).

	Logatoms		Words		Sentences	
	Machine intelligibility	Correlation	Machine intelligibility	Correlation	Machine intelligibility	Correlation
Small	30.6	.81*** (19.7)	36.7	.77*** (24.0)	32.2	.71*** (31.6)
Medium	27.1	.83*** (18.9)	25.0	.83*** (23.7)	20.5	.73*** (31.4)
Large	4.2	.83*** (22.8)	12.5	.76*** (25.1)	15.8	.70*** (32.0)

Note. The root-mean-square error obtained using a linear regression is reported between brackets.
***$p < .001$; one-tailed test.

behavioral scores: 62.5% versus 8.8% for logatoms, 85.0% versus 70.0% for words, and 79.8% versus 46.5% for sentences (see the corresponding data points in the lower row of Figure 3).

Increasing the lexicon size to medium and large resulted in a progressive reduction in machine intelligibility, independently of the speech material used. The quality of the prediction of human intelligibility scores was assessed in terms of both the strength of the association between human and machine intelligibility scores and prediction errors. As human and machine scores were not normally distributed, the strength of

association between both variables, and its statistical significance, were assessed through Spearman's correlations. To estimate prediction errors, simple parametric linear regressions were used to calculate root-mean-square errors (RMSEs). The results show that, contrary to the ASR performance, the quality of the predictions, either in terms of Spearman's *rho* or RMSE, did not systematically change as a function of lexicon size. Given the lesser processing costs and shorter processing time associated with the use of an ASR system implementing a small lexicon, further analyses were limited to machine intelligibility obtained with the small lexicon.

The lower row of Figure 3 shows scatterplots relating human intelligibility scores to machine intelligibility scores. For logatoms (left panel), the data are roughly following the diagonal, showing an audibility gradient from "high PTA" (yielding both low human and low machine intelligibility scores) to "low PTA" (yielding both high human and high machine intelligibility scores). Such a linear relationship between human and machine scores is not observed with the other two speech materials affording the use of linguistic top-down information (Pichora-Fuller, 2008): For OHI listeners with PTAs of 40 dB HL and above (excluding the outlier mentioned earlier), predicted intelligibility remains low (i.e., does not exceed 53.4%) in the face of human performance varying widely from 0.0 to 81.7% for words and from 0.0 to 95.0% for sentences.

As done in Fontan et al. (2017), ARHL was initially simulated by spectral attenuation, spectral smearing, and expansion of the signal envelope to a power, mimicking three effects of ARHL: HT elevation, FS loss, and LR. To quantify the relative contribution of these different components, the present study also evaluated machine intelligibility for speech signals processed to mimic only one or a combination of the components of ARHL (see Table 3).

Surprisingly, the HT-elevation condition yielded Spearman correlation coefficients that were (marginally) higher than those for the condition in which all three effects of ARHL were simulated simultaneously (HT elevation + FS loss + LR). Adding FS loss to HT elevation resulted in small decreases in the strength of the correlation for logatoms and words (of .07 and .06 respectively) and an even smaller increase for sentences (of .02). Associated RMSEs all increased (by 0.6, 2.4, and 1.7 for logatoms, words, and sentences, respectively). On the other hand, removing FS loss from the full ARHL simulation improved the strength of the correlations for all three speech materials (by .08, .05, and .09 for logatoms, words, and sentences, respectively) and reduced the RMSE (by 1.6, 2.2, and 2.2 for logatoms, words, and sentences, respectively). After applying a Fisher's *r*-to-*Z* transformation (Lee & Preacher, 2013), the improvements in the strength of the correlations were found to be significant for logatoms ($Z = 2.12$, $p = .017$, one-tailed) and for sentences ($Z = 2.90$, $p = .004$, one-tailed). When adding the simulation of LR to the simulation of HT elevation, or to the simulation of HT elevation + FS loss, both increases and decreases in the strength of the correlations and in the RMSEs were observed. However, these changes were significant only for the simulation of HT elevation with logatoms ($Z = 2.18$, $p = .015$, one-tailed) and sentences ($Z = 1.89$, $p = .029$, one-tailed), for which the correlation coefficient increased by .07 and .05, respectively.

Table 3. Spearman Correlation Coefficients Between Human Intelligibility Scores and Machine Intelligibility Scores (Obtained With the Small Lexicon) for Each of the Three Speech Materials, as a Function of ARHL-Simulation Condition (HT elevation, HT elevation + FS loss, HT elevation + LR, and HT elevation + FS loss + LR).

	Human intelligibility		
ARHL simulation	Logatoms	Words	Sentences
HT	.82*** (19.0)	.86*** (21.7)	.75*** (29.6)
HT + FS	.75*** (19.6)	.80*** (24.1)	.77*** (31.3)
HT + LR	.89*** (18.1)	.84*** (21.8)	.80*** (29.4)
HT + FS + LR	.81*** (19.7)	.79*** (24.0)	.71*** (31.6)

Note. The root-mean-square error obtained using a linear regression is reported between brackets. ARHL = age-related hearing loss. HT = elevation of hearing thresholds; FS = loss of frequency selectivity; LR = loudness recruitment.
***$p < .001$; one-tailed test.

Discussion

This study aimed at predicting the speech-identification performance in quiet of unaided OHI listeners for speech materials of varying linguistic complexity and predictability that are commonly used in audiological practice in France (Rembaud et al., 2017). Twenty-four OHI listeners completed three speech-identification tasks consisting in the repetition of logatoms, words, and sentences. An ASR system was used to recognize the same speech stimuli, which were processed to mimic some of the perceptual consequences experienced by each listener (i.e., elevation of HT, loss of FS, and LR).

Human and Machine Speech-Intelligibility Performance

Despite all speech materials being pronounced by the same speaker and recorded and reproduced under the same acoustic conditions, human speech-intelligibility performance improved from logatoms over words to sentences, that is, with increasing linguistic context (i.e., lexical, morphosyntactic, and semantic information). This observation is in line with the Mutuality Model of

Lindblom (1990), according to which listeners will take advantage of nonacoustic cues when the speech signal is degraded. Also, the psychometric functions relating human intelligibility scores to mean audibility shifted toward higher degrees of signal degradation, associated with higher PTAs, for words and sentences. Similar observations have been made for normal-hearing and HI listeners performing identification-in-noise tasks, using speech materials of varying predictability (Fontan, Tardieu, et al., 2015; Kryter, 1994; Pichora-Fuller, 2008).

By comparison, machine intelligibility did not markedly or consistently improve with the speech materials. The discrepancies observed between human and machine intelligibility could be due to the fact that the acoustic models used in the present study were not trained on degraded speech (in contrast to those used in Kollmeier et al., 2016; Schädler et al., 2018). One way to improve machine intelligibility scores would therefore be to train acoustic models on speech material that has previously been processed to simulate HT elevation, FS loss, and LR of various degrees. A second (less costly, but probably also less efficient) way to achieve more robust acoustic models (i.e., models that are less sensible to degraded conditions) would be to modify the acoustic models by using adaptation techniques such as maximum likelihood linear regression (Leggetter & Woodland, 1995) or maximum a posteriori (Gauvain & Lee, 1994) on a (smaller) corpus of speech processed to mimic ARHL.

The size of the lexicon had an impact on machine performance: The smaller the lexicon, the higher the recognition scores. This was expected, as there are fewer alternative candidates for speech recognition in case of a smaller lexicon, and, thus, there is a greater probability for the ASR system to successfully recognize the target utterance. Given the higher performance and lower associated processing costs, only small lexicons were considered for the prediction of human intelligibility scores. It should however be noted that, if aiming at the *qualitative* prediction of human speech-identification performance (such as phonemic confusions; Fontan et al., 2016), using small lexicons might not be optimal. Indeed, as the lexicon determines to a large extend the errors and confusions the ASR system can perform, a qualitative speech-intelligibility prediction system should incorporate a lexicon that matches as closely as possible the mental lexicon of the human listener.

Prediction of Human Speech-Intelligibility Scores

Consistent with previous studies (Fontan et al., 2017; Kollmeier et al., 2016; Schädler et al., 2018), very strong and highly significant correlations were observed between human and machine intelligibility scores. Importantly, this result was achieved for the first time using an ASR system that was trained on unprocessed speech material (i.e., not degraded to mimic the consequences of ARHL) and different from the test material.

However, for all speech materials and ARHL-simulation conditions, the observed RMSEs were very large, indicating that the prediction system could mainly predict *trends* in human speech intelligibility. Among the three speech materials used in the present study, the strongest correlation (Spearman's $rho = .89$) and the lowest RMSE (18.1%) were observed for logatoms. This is likely due to the prediction system mainly reflecting the functioning of the peripheral part of the auditory system. Indeed, the ASR system is basically an acoustic-phonetic decoder: Aside from syntactic information (i.e., information on which word may occur after a given word), only bottom-up, acoustic-phonetic information is processed by the system. Because human performance on the logatom-identification task is also mainly determined by acoustic-phonetic information, a very high correlation was found between human and machine intelligibility scores for this speech material.

In contrast to human listeners, the ASR system could not compensate for the degradation of the speech signal by taking advantage of the linguistic information present in the case of words and sentences. This resulted in an increasing discrepancy between human and machine intelligibility scores with increasing linguistic context.

Also, as cognitive abilities contribute to some extent to speech processing in quiet in unaided OHI listeners (van Rooij & Plomp, 1992), variability in these abilities across the listeners (that are not echoed in the ASR system) most likely reduced the association between human and machine intelligibility scores. To explore whether general cognitive functioning also contributed to speech intelligibility in our sample, bivariate correlations between MMSE scores and human intelligibility scores were calculated. While the correlations failed to be significant for logatoms (Spearman's $rho = .33$, $p = .059$, one-tailed) and for words (Spearman's $rho = .30$, $p = .075$, one-tailed), the correlation was found to be significant for sentences (Spearman's $rho = .40$, $p = .026$, one-tailed). Also, separate multiple linear regressions were conducted for each of the speech materials, with human intelligibility scores as the dependent variable and machine intelligibility scores (obtained using the small lexicon and the AHRL-simulation condition yielding the strongest correlation with human intelligibility scores) and MMSE scores as predictor variables. For each speech material, MMSE scores were always entered after machine intelligibility scores. Results are shown in Table 4 and indicate that the addition of MMSE scores just failed to contribute significantly to the prediction of human intelligibility scores for logatoms ($p = .074$) and words ($p = .056$) but significantly improved the prediction for sentences ($p = .024$). The normality assumption for the three models was assessed by a Kolmogorov–

Table 4. Results of Multiple Linear Regressions for the Prediction of Human Logatom-, Word-, and Sentence-Intelligibility Scores Using Machine Intelligibility Scores and MMSE Scores as Predictor Variables.

Speech material	R^2	RMSE	Predictor	β coef	R^2 change	p value
Logatoms	.78	17.2	Machine intelligibility	.828	.739	<.001
			MMSE score	.196	.038	.074
Words	.65	21.7	Machine intelligibility	.750	.584	<.001
			MMSE score	.261	.068	.056
Sentences	.52	26.6	Machine intelligibility	.602	.387	.001
			MMSE score	.367	.134	.024

Note. Machine intelligibility scores were always entered first into the regression analysis. For each of the three speech materials, the explained variance (R^2) and root-mean-square error (RMSE) for the entire model are given, as well as the contribution of each predictor in terms of its standardized coefficient (β coef), associated change in the amount of explained variance (R^2 change), and significance (p value). MMSE = Mini-Mental State Examination.

Smirnov test that indicated that the distributions of the prediction residuals were not significantly different from a normal distribution (all $p = .200$). The amount of additional variance explained by the MMSE scores increased as a function of the linguistic complexity of the speech material: It was 3.8, 6.8, and 13.4 percentage points for logatoms, words, and sentences, respectively.

It is noteworthy that all our participants had MMSE scores ≥ 27 (out of 30), indicative of normal cognitive functioning, and that the MMSE has poor discriminative power in the nonpathological range of cognitive functioning. Hence, it can be speculated that the use of more sensitive cognitive tests and a more cognitively heterogeneous sample of OHI listeners would reveal a greater potential for cognition to improve ASR-based predictions of speech intelligibility.

The current study also investigated the additional effects of simulated FS loss and LR on the prediction of human speech intelligibility, when these were added separately or together to the simulation of HT elevation. One possible concern was that the spectral smearing used to simulate FS loss might have no effect on ASR performance because the computation of MFCC features also involves spectral smearing. It was thus possible that the simulation of FS loss would remain "invisible" in the ASR features unless the amount of smearing used to simulate FS loss would exceed that imposed by the computation of MFCCs. However, in the present study and in the study of Fontan et al. (2017), the simulation of FS loss did affect ASR performance, and this was observed for all levels of smearing used (i.e., mild, moderate, and severe; results not shown). Contrary to the simulation of LR which yielded marginal improvements in the correlation between human and machine intelligibility scores and, in some cases, in RMSEs, the inclusion of the simulation of FS loss yielded almost always weaker correlations and always yielded higher RMSEs. One explanation might be that artifacts produced by the spectral smearing to mimic the loss of FS had detrimental effects on machine intelligibility; such signal-processing-related artifacts do not occur in the impaired human ear. To overcome this issue, future studies could train the acoustic models of the ASR on speech signals processed through the ARHL simulator; in this case, great care should be taken in considering which ARHL-simulation conditions (i.e., which ARHL profiles) should be selected for the training of the acoustic models, as it may result in an overfitting of the system (i.e., to higher ASR scores) for OHI listeners whose audiometric profiles are close to the training conditions. To avoid such a bias, a separate set of acoustic models could be created for each listener's audiometric profile. In that case, all the listeners would benefit from the same amount of acoustic training. However, given the large training speech corpus, resulting in a computation time of more than 30 hr for a single set of acoustic models, this approach was deemed beyond the scope of the present study.

Another limit of the prediction system used in this study is that the strength of the simulation of FS loss and LR depended on the category of the severity of the audiometric loss (mild, moderate, or severe). For listeners whose PTA falls just below or above the threshold associated with a given ARHL severity category, this could result in an under- or oversimulation of FS loss and LR. This is possibly illustrated by the three outliers in the bottom panels of Figure 3, representing data for the same listener with a PTA of 55.2 dB HL. As this PTA just falls short of the lower limit of the severe-ARHL category (i.e., 56 dB HL), the simulated FS loss and LR probably underestimated the actual consequences of ARHL in this individual. This would have resulted in higher machine than human intelligibility scores, as was empirically observed for this listener.

Also, FS loss and LR show a high interindividual variability that is unrelated to audiometric thresholds (Al-Salim et al., 2010; Hopkins & Moore, 2011; Marozeau & Florentine, 2007). Thus, individual FS loss and LR might have to be measured experimentally, rather than being inferred, to improve the prediction of human intelligibility scores. To test this hypothesis, LR was estimated for each participant by computing the

average of the differences between HTs and the maximum comfortable levels at all audiometric frequencies. Separate multiple linear regressions were computed for each speech material, with human intelligibility scores as the dependent variable and machine intelligibility scores and LR estimates as the predictor variables. The results indicate that LR estimates do not significantly improve the prediction of the human intelligibility scores for any of the speech materials (all $p \geq .097$). This finding might indicate that the ARHL simulation successfully approximated the individual levels of LR experienced by the listeners. On the other hand, the fact that the average dynamic range across frequencies did not contribute significantly to the model could indicate that it is not a good estimate of LR. A number of alternative methods for assessing LR exist, such as asking the listeners to estimate the intensity of pure tones with absolute numbers (e.g., Hellman & Meiselman, 1990) or on a categorical scale (e.g., Brand & Hohmann, 2001). Similarly, collecting individual data on FS loss (through the assessment of auditory filter bandwidth) could help to better simulate ARHL and therefore achieve better predictions of human intelligibility.

The prediction system in this study also did not take into account changes in the sensitivity to temporal-envelope and temporal-fine-structure information that occurs with age (e.g., Füllgrabe et al., 2018; Moore et al., 2012) and hearing loss (e.g., Füllgrabe & Moore, 2017; Gallun et al., 2014; King et al., 2014) and that are associated with speech-in-noise perception (e.g., Füllgrabe et al., 2015; Lopez-Poveda et al., 2017; Strelcyk & Dau, 2009). To our knowledge, there is currently no generally accepted simulation of the reduction in or loss of temporal processing abilities that could be used in addition to the ARHL simulation used here. However, ASR-based predictions of human speech intelligibility could possibly be improved by statistically taking into account temporal processing abilities that would need to be measured for each listener.

Finally, the present study investigated speech intelligibility in conditions that are not the most representative of real-life listening. First, conversational speech levels are generally higher than the presentation level used in the present study. A lower presentation level of 50 dB SPL was used, and, as the ASR system uses intensity normalization, predictions would most likely result in an underestimation of human performance at higher presentation levels. Second, the present study focused on the prediction of speech intelligibility in quiet by unaided OHI listeners. However, the main difficulty for people with ARHL is to understand speech in noisy environments, and this is observed even when amplification is provided by HAs. Thus, the suitability of ASR for the prediction of unaided and aided speech-in-noise perception needs to be assessed in future studies. This would require the training of acoustic models to cope with the consequences of signal processing schemes used in HAs (e.g., linear amplification, amplitude compression, frequency shifts) and the presence of different types of background noise (e.g., white noise, speech babble).

Conclusions

This study compared logatom-, word-, and sentence-identification performance of 24 unaided OHI listeners with predictions of an ASR system presented with the same speech stimuli but processed to simulate some of the consequences of ARHL (elevation of HT, loss of FS, and LR). Strong to very strong correlations were observed between human and machine intelligibility scores for all three speech materials, but in all cases, RMSEs were large. Simulating FS loss, in addition to the elevation of HT, resulted in weaker correlations and higher RMSEs. The strongest correlations were observed for logatoms for which both human and machine performance rely on acoustic cues. Correlations were weaker for words and sentences for which the ASR system did not benefit from the additional linguistic context present in those speech materials. The prediction of human sentence-identification performance was significantly improved by taking into account general cognitive ability of the listener.

Acknowledgments

The authors thank Etienne Revol-Buisson from the cabinet d'audioprothèse "Espace Audition" in Montauban (France) and Drs. Nathaniel Khalifa and Vincent Calas from the Honoré Cave Hospital in Montauban (France) for their help with participant recruitment. The authors are grateful to Prof. Brian C. J. Moore and Dr. Michael A. Stone for providing the ARHL-simulation program and Dr. Maxime Le Coz for helping with signal processing and the statistical analyses. The authors also thank the Associate Editor Dr. Torsten Dau, three anonymous reviewers, and Dr. Tom Baer for their comments on previous versions of the article.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: This research was conducted as part of product/service development at Archean Technologies (Montauban, France), who filed a European patent describing the use of automatic speech recognition for measuring the performance of sound distribution systems (Aumont & Wilhem-Jaureguiberry, 2009). At the time of the research project, the last author acted as a scientific consultant for Archean Technologies.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The project received support from the Institute of Advanced Studies of Loughborough University (UK) in the form of a visiting fellowship to Dr. Lionel Fontan.

ORCID iDs

Lionel Fontan https://orcid.org/0000-0001-7895-8567
Christian Füllgrabe https://orcid.org/0000-0001-9127-8136

References

Al-Salim, S. C., Kopun, J. G., Neely, S. T., Jesteadt, W., Stiegemann, B., & Gorga, M. P. (2010). Reliability of categorical loudness scaling and its relation to threshold. *Ear and Hearing*, *31*(4), 567–578. https://doi.org/10.1097/AUD.0b013e3181da4d15

American National Standard Institute. (1969). *American National Standard specification for audiometers*.

American National Standard Institute. (1997). *Methods for the calculation of the speech intelligibility index*.

Aumont, X., & Wilhem-Jaureguiberry, A. (2009). *European Patent No. 2136359 – Method and Device for Measuring the Intelligibility of a Sound Distribution System*. Courbevoie, France: Institut National de la Propriété Industrielle.

Baer, T., & Moore, B. C. J. (1993). Effects of spectral smearing on the intelligibility of sentences in noise. *Journal of the Acoustical Society of America*, *94*(3), 1229–1241. https://doi.org/10.1121/1.408176

Baltes, P. B., & Lindenberger, U. (1997). Emergence of a powerful connection between sensory and cognitive functions across the adult life span: A new window to the study of cognitive aging? *Psychology and Aging*, *12*(1), 12–21. https://doi.org/10.1037/0882-7974.12.1.12

Barker, J., & Cooke, M. (2007). Modelling speaker intelligibility in noise. *Speech Communication*, *49*(5), 402–417. https://doi.org/10.1016/j.specom.2006.11.003

Bassuk, S. S., Wypij, D., & Berkman, L. F. (2000). Cognitive impairment and mortality in the community-dwelling elderly. *American Journal of Epidemiology*, *151*(7), 676–688. https://doi.org/10.1093/oxfordjournals.aje.a010262

Brand, T., & Hohmann, V. (2001). Effect of hearing loss, centre frequency, and bandwidth on the shape of loudness functions in categorical loudness scaling. *Audiology*, *40*, 92–103. https://doi.org/10.3109/00206090109073104

British Society of Audiology. (2011). *Recommended procedure. Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking*. http://www.thebsa.org.uk/wp-content/uploads/2014/04/BSA_RP_PTA_FINAL_24Sept11_MinorAmend06Feb12.pdf

Bruce, M. L., Hoff, R. A., Jacobs, S. C., & Leaf, P. J. (1995). The effects of cognitive impairment on 9-year mortality in a community sample. *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *50*(6), 289–296. https://doi.org/10.1093/geronb/50B.6.P289

Carroll, R., Warzybok, A., Kollmeier, B., & Ruigendijk, E. (2016). Age-related differences in lexical access relate to speech recognition in noise. *Frontiers in Psychology*, *7*, 990. https://doi.org/10.3389/fpsyg.2016.00990

Collège National d'Audioprothèse. (2007). *Précis d'audioprothèse—Tome II* [Reference manual for hearing-aid specialists—Part II]. Elsevier-Masson.

Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, B. E., Klein, R., Mares-Perlman, J. A., & Nondahl, D. M. (1998). Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin: The epidemiology of hearing loss study. *American Journal of Epidemiology*, *148*(9), 879–886. https://doi.org/10.1093/oxfordjournals.aje.a009713

Crum, R. M., Anthony, J. C., Bassett, S. S., & Folstein, M. F. (1993). Population-based norms for the mini-mental state examination by age and educational level. *JAMA*, *269*(18), 2386–2391. https://doi.org/10.1001/jama.1993.03500180078038

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

de Calmès, M., Farinas, J., Ferrané, I., & Pinquier, J. (2005, January). *Campagne ESTER : Une première version d'un système complet de transcription automatique de la parole grand vocabulaire* [ESTER campaign: A first version of a complete automatic speech transcription system with a large vocabulary] [Paper presentation]. Atelier ESTER, Avignon, France.

Deléglise, P., Esteve, Y., Meignier, S., & Merlin, T. (2005, September). *The LIUM speech transcription system: A CMU Sphinx III-based system for French broadcast news* [Paper presentation]. Proceedings of Interspeech '05, Lisbon, Portugal.

Dodelé, L., & Dodelé, D. (2000). L'audiométrie vocale en présence de bruit et le test AVfB [Speech-in-noise audiometry and the AVfB test]. *Les Cahiers de l'Audition*, *3*(6), 15–22.

Estève, Y. (2009). *Traitement automatique de la parole : Contributions* [Automatic speech processing: Contributions] (Habilitation à diriger les recherches). Université du Maine.

Falk, T. H., Parsa, V., Santos, J. F., Arehart, K., Hazrati, O., Huber, R., . . . Scollie, S. (2015). Objective quality and intelligibility prediction for users of assistive listening devices. *IEEE Signal Processing Magazine*, *32*(2), 114–124. https://doi.org/10.1109/MSP.2014.2358871

Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., & Aumont, X. (2016). Using phonologically weighted Levenshtein distances for the prediction of microscopic intelligibility. In *Proceedings of Interspeech '16* (pp. 650–654). The International Speech and Communication Association (ISCA). https://doi.org/10.21437/Interspeech.2016-431

Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., Tardieu, J., Magnen, C., . . . Füllgrabe, C. (2017). Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language, and Hearing Research*, *60*(9), 2394–2405. https://doi.org/10.1044/2017_JSLHR-S-16-0269

Fontan, L., Magnen, C., Tardieu, J., Ferrané, I., Pinquier, J., Farinas, J., . . . Aumont, X. (2014). Comparaison de mesures perceptives et automatiques de l'intelligibilité de la parole : Cas de la parole dégradée par une simulation de la presbyacousie [Comparison of perceptive and automatic measures of intelligibility: Application to speech simulating age-related hearing loss]. *Traitement Automatique des Langues*, *55*(2), 151–174.

Fontan, L., Pellegrini, T., Olcoz, J., & Abad, A. (2015). Predicting disordered speech comprehensibility from goodness of pronunciation scores. In *Proceedings of the Sixth Workshop on Speech and Language Processing for Assistive Technologies: SLPAT 2015 – Satellite Workshop of Interspeech '15*. The International Speech and Communication Association (ISCA). http://www.slpat.org/slpat2015/papers/fontan-pellegrini-olcoz-abad.pdf

Fontan, L., Tardieu, J., Gaillard, P., Woisard, V., & Ruiz, R. (2015). Relationship between speech intelligibility and speech comprehension in babble noise. *Journal of Speech, Language, and Hearing Research*, *58*(3), 977–986. https://doi.org/10.1044/2015_jslhr-h-13- 0335

Fournier, J. E. (1951). *Audiométrie vocale : Les epreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités* [Speech audiometry: Speech-intelligibility tests and their application for the diagnosis, survey and hearing-aid rehabilitation of hearing losses]. Maloine.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, *19*(1), 90–119. https://doi.org/10.1121/1.1916407

Füllgrabe, C. (2013). Age-dependent changes in temporal-fine-structure processing in the absence of peripheral hearing loss. *American Journal of Audiology*, *22*(2), 313–315. https://doi.org/10.1044/1059-0889(2013/12-0070)

Füllgrabe, C., Harland, A. J., Sęk, A. P., & Moore, B. C. J. (2017). Development of a method for determining binaural sensitivity to temporal fine structure. *International Journal of Audiology*, *56*(12), 926–935. https://doi.org/10.1080/14992027.2017.1366078

Füllgrabe, C., Meyer, B., & Lorenzi, C. (2003). Effect of cochlear damage on the detection of complex temporal envelopes. *Hearing Research*, *178*(1–2), 35–43. https://doi.org/10.1016/S0378-5955(03)00027-3

Füllgrabe, C., & Moore, B. C. J. (2017). Evaluation of a method for determining binaural sensitivity to temporal fine structure (TFS-AF test) for older listeners with normal and impaired low-frequency hearing. *Trends in Hearing*, *21*, 233121651773723. https://doi.org/10.1177/2331216517737230

Füllgrabe, C., & Moore, B. C. J. (2018). The association between the processing of binaural temporal-fine-structure information and audiometric threshold and age: A meta-analysis. *Trends in Hearing*, *22*, 233121651879725. https://doi.org/10.1177/2331216518797259

Füllgrabe, C., Moore, B. C. J., & Stone, M. A. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience*, *6*, 347. https://doi.org/10.3389/fnagi.2014.00347

Füllgrabe, C., & Rosen, S. (2016a). Investigating the role of working memory in speech-in-noise identification for listeners with normal hearing. *Advances in Experimental Medicine and Biology*, 29–36. https://doi.org/10.1007/978-3-319-25474-6_4

Füllgrabe, C., & Rosen, S. (2016b). On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds. *Frontiers in Psychology*, *7*, 1268. https://doi.org/10.3389/fpsyg.2016.01268

Füllgrabe, C., Sęk, A. P., & Moore, B. C. J. (2018). Senescent changes in sensitivity to binaural temporal fine structure. *Trends in Hearing*, *22*, 2331216518788224. https://doi.org/10.1177/2331216518788224

Galliano, S., Gravier, G., & Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of Interspeech '09*. The International Speech and Communication Association (ISCA). https://www.irisa.fr/metiss/ggravier/biblio/09/galliano-interspeech-09.pdf

Gallun, F. J., McMillan, G. P., Molis, M. R., Kampel, S. D., Dann, S. M., & Konrad-Martin, D. L. (2014). Relating age and hearing loss to monaural, bilateral, and binaural temporal sensitivity. *Frontiers in Neuroscience*, *8*, 172. https://doi.org/10.3389/fnins.2014.00172

Gauvain, J.-L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 291–298. https://doi.org/10.1109/89.279278

Glasberg, B. R., & Moore, B. C. J. (1986). Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *Journal of the Acoustical Society of America*, *79*(4), 1020–1033. https://doi.org/10.1121/1.393374

Grose, J. H., & Mamo, S. K. (2010). Processing of temporal fine structure as a function of age. *Ear and Hearing*, *31*(6), 755–760. https://doi.org/10.1097/AUD.0b013e3181e627e7

He, N., Mills, J. H., Ahlstrom, J. B., & Dubno, J. R. (2008). Age-related differences in the temporal modulation transfer function with pure-tone carriers. *Journal of the Acoustical Society of America*, *124*(6), 3841–3849. https://doi.org/10.1121/1.2998779

Hellman, R. P., & Meiselman, C. H. (1990). Loudness relations for individuals and groups in normal and impaired hearing. *Journal of the Acoustical Society of America*, *88*, 2596–2606. https://doi.org/10.1121/1.399979

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., & Benson, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders*, *17*(3), 321–337. https://doi.org/10.1044/jshd.1703.321

Hopkins, K., & Moore, B. C. J. (2011). The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise. *Journal of the Acoustical Society of America*, *130*(1), 334–349. https://doi.org/10.1121/1.3585848

Hudgins, C. V., Hawkins, J., Kaklin, J., & Stevens, S. (1947). The development of recorded auditory tests for measuring

hearing loss for speech. *Laryngoscope*, *57*(1), 57–89. https://doi.org/10.1288/00005537-194701000-00005

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*, *12*(2), 198–208. https://doi.org/10.1044/1058-0360(2003/066)

Kalafat, M., Hugonot-Diener, L., & Poitrenaud, J. (2003). Standardisation et étalonnage français du Mini Mental State (MMS) version GRÉCO [The Mini Mental State (MMS): French standardization and normative data]. *Revue de Neuropsychologie*, *13*(2), 209–236.

King, A., Hopkins, K., & Plack, C. J. (2014). The effects of age and hearing loss on interaural phase difference discrimination. *Journal of the Acoustical Society of America*, *135*(1), 342–351. https://doi.org/10.1121/1.4838995

Kollmeier, B., Schädler, M. R., Warzybok, A., Meyer, B. T., & Brand, T. (2016). Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with FADE: Empowering the attenuation and distortion concept by Plomp with a quantitative processing model. *Trends in Hearing*, *20*, 233121651665579. https://doi.org/10.1177/2331216516655795

Kryter, K. (1994). *The handbook of hearing and the effects of noise: Physiology, psychology, and public health*. Academic Press.

Leduc, R. (1997). *Pour la réussite du dépistage précoce et continu* [For a successful early and continuous screening]. Centre Franco-Ontarien de Ressources Pédagogiques.

Lee, I. A., & Preacher, K. J. (2013). *Calculation for the test of the difference between two dependent correlations with one variable in common*. http://quantpsy.org

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, *9*(2), 171–185. https://doi.org/10.1006/csla.1995.0010

Lindblom, B. (1990). On the communication process: Speaker-listener interaction and the development of speech. *Augmentative and Alternative Communication*, *6*, 220–230. https://doi.org/10.1080/07434619012331275504

Lopez-Poveda, E. A., Johannesen, P. T., Pérez-González, P., Blanco, J. L., Kalluri, S., & Edwards, B. (2017). Predictors of hearing-aid outcomes. *Trends in Hearing*, *21*, 233121651773052. https://doi.org/10.1177/2331216517730526

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., & Nöth, E. (2009). PEAKS – A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, *51*(5), 425–437. https://doi.org/10.1016/j.specom.2009.01.004

Marozeau, J., & Florentine, M. (2007). Loudness growth in individual listeners with hearing losses: A review. *Journal of the Acoustical Society of America*, *122*(3), EL81 -EL87. doi: 10.1121/1.2761924

Moore, B. C. J., & Glasberg, B. R. (1993). Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech. *Journal of the Acoustical Society of America*, *94*(4), 2050–2062. https://doi.org/10.1121/1.407478

Moore, B. C. J. (2007). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues* (2nd ed.). Chichester, England: Wiley. doi: 10.1002/9780470987889

Moore, B. C. J., Glasberg, B. R., Stoev, M., Füllgrabe, C., & Hopkins, K. (2012). The influence of age and high-frequency hearing loss on sensitivity to temporal fine structure at low frequencies (L). *Journal of the Acoustical Society of America*, *131*(2), 1003–1006. https://doi.org/10.1121/1.3672808

Moulin, A., Bernard, A., Tordella, L., Vergne, J., Gisbert, A., Martin, C., & Richard, C. (2016). Variability of word discrimination scores in clinical practice and consequences on their sensitivity to hearing loss. *European Archives of Oto-Rhino-Laryngology, 274*, 2117–2124. https://doi.org/10.1007/s00405-016-4439-x

Neher, T., Laugesen, S., Søgaard Jensen, N., & Kragelund, L. (2011). Can basic auditory and cognitive measures predict hearing-impaired listeners' localization and spatial speech recognition abilities? *Journal of the Acoustical Society of America*, *130*(3), 1542–1558. https://doi.org/10.1121/1.3608122

Nejime, Y., & Moore, B. C. J. (1997). Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise. *Journal of the Acoustical Society of America*, *102*(1), 603–615. https://doi.org/10.1121/1.419733

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661–677. https://doi.org/10.1017/S014271640707035X

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Pichora-Fuller, K. (2008). Use of supportive context by younger and older adult listeners: Balancing bottom-up and top-down information processing. *International Journal of Audiology*, *47*(Suppl 2), S72–S82. https://doi.org/10.1080/14992020802307404

Plomp, R. (1978). Auditory handicap of hearing impairment and the limited benefit of hearing aids. *Journal of the Acoustical Society of America*, *63*(2), 533–549. https://doi.org/10.1121/1.381753

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*, 257–286. https://doi.org/10.1109/5.18626

Rembaud, F., Fontan, L., & Füllgrabe, C. (2017). L'audiométrie vocale en France : État des lieux [Speech audiometry in France: Current clinical practices]. *Cahiers de l'Audition*, *6*, 22–25.

République Française. (2017). *Code de la santé publique. Audioprothésiste—Local réservé à l'activité professionnelle* [Public health code. Hearing-aid dispenser—Professional premises]. https://www.legifrance.gouv.fr/

Ross, B., Fujioka, T., Tremblay, K. L., & Picton, T. W. (2007). Aging in binaural hearing begins in mid-life: Evidence from cortical auditory-evoked responses to changes in interaural phase. *Journal of Neuroscience*, *27*(42), 11172–11178. https://doi.org/10.1523/JNEUROSCI.1813-07.2007

Ryu, I. S., Ahn, J. H., Lim, H. W., Joo, K. Y., & Chung, J. W. (2012). Evaluation of masking effects on speech perception in patients with unilateral chronic tinnitus using the hearing in noise test. *Otology & Neurotology*, *33*(9), 1472–1476. https://doi.org/10.1097/MAO.0b013e31826dbcc4

Schädler, M. R., Warzybok, A., Hochmuth, S., & Kollmeier, B. (2015). Matrix sentence intelligibility prediction using an automatic speech recognition system. *International Journal of Audiology*, *54*(Suppl 2), 100–107. https://doi.org/10.3109/14992027.2015.1061708

Schädler, M. R., Warzybok, A., & Kollmeier, B. (2018). Objective prediction of hearing aid benefit across listener groups using machine learning: Speech recognition performance with binaural noise-reduction algorithms. *Trends in Hearing*, *22*, 233121651876895. https://doi.org/10.1177/2331216518768954

Seymore, K., Chen, S., Doh, S., Eskenazi, M., Gouvêa, E., Raj, B., . . . Thayer, E. (1998). *The 1997 CMU Sphinx-3 English broadcast news transcription system* [Paper presentation]. Proceedings of the 1998 DARPA Speech Recognition Workshop, Lansdowne, Canada.

Sheldon, S., Pichora-Fuller, M. K., & Schneider, B. A. (2008). Priming and sentence context support listening to noise-vocoded speech by younger and older adults. *Journal of the Acoustical Society of America*, *123*(1), 489–499. https://doi.org/10.1121/1.2783762

Spille, C., Ewert, S. D., Kollmeier, B., & Meyer, B. T. (2018). Predicting speech intelligibility with deep neural networks. *Computer Speech & Language*, *48*, 51–66. https://doi.org/10.1016/j.csl.2017.10.004

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America*, *67*(1), 318–326. https://doi.org/10.1121/1.384464

Stone, M. A., Moore, B. C. J., Füllgrabe, C., & Hinton, A. C. (2009). Multichannel fast-acting dynamic range compression hinders performance by young, normal-hearing listeners in a two-talker separation task. *Journal of the Audio Engineering Society*, *57*(7/8), 532–546.

Strelcyk, O., & Dau, T. (2009). Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *Journal of the Acoustical Society of America*, *125*(5), 3328–3345. https://doi.org/10.1121/1.3097469

Tyler, R. S., Hall, J. W., Glasberg, B. R., Moore, B. C. J., & Patterson, R. D. (1984). Auditory filter asymmetry in the hearing impaired. *Journal of the Acoustical Society of America*, *76*(5), 1363–1368. https://doi.org/10.1121/1.391452

Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A., . . . Giguère, C. (2005). Adaptation of the HINT (hearing in noise test) for adult Canadian francophone populations. *International Journal of Audiology*, *44*(6), 358–361. https://doi.org/10.1080/14992020500060875

van Rooij, J. C. G. M., & Plomp, R. (1992). Auditive and cognitive factors in speech perception by elderly listeners. III. Additional data and final discussion. *Journal of the Acoustical Society of America*, *91*(2), 1028–1033. https://doi.org/10.1121/1.402628

Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age–cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin*, *122*(3), 231–249. https://doi.org/10.1037/0033-2909.122.3.231

Zaudig, M. (1992). A new systematic method of measurement and diagnosis of "mild cognitive impairment" and dementia according to ICD-10 and DSM-III-R criteria. *International Psychogeriatrics*, *4*(Suppl 2), 203–219. https://doi.org/10.1017/S1041610292001273