Turning up the Background Noise; Optimizing Intelligibility and Comprehension of Broadcast Material for Hard of Hearing Listeners

Lauren Ward[1]

[1] Acoustics Research Centre, University of Salford, Manchester, UK
L.Ward7@edu.salford.ac.uk

1. Introduction

The *'cocktail party problem'*, coined by Cherry in the 1950s, refers to the challenge of finding speech intelligible in noisy or complex acoustic scenes [1]. This problem is even more profound for the 11 million UK individuals who have some degree of hearing impairment [2, 3], 91.7% of whom have mild to moderate loss [4]. However such complex acoustic scenes are pervasive, even in broadcast content where presenters are often masked by wind noise or a commentator's voice competes with a cheering crowd [5]. Studies have shown that as many as 87% of hearing impaired listeners struggle to understand speech in broadcast material [6], with the effects of background noise on understanding being a core complaint [7].

In response, an increasing body of work has focused on the creation of *'clean audio'*; defined as broadcast audio providing improved intelligibility targeted for listeners with hearing impairments [8]. The provision of a clean audio option, in complement to subtitling, is motivated by the fact that not only Hard of Hearing (HoH) but also normal hearing listeners report the adverse affect of background sounds (45% in a Royal National Institute for the Deaf study [9]). Further to this, it has been shown that listeners with audiometrically normal pure tone thresholds differ in their ability to attend to speech in noise [10]. This means even those with nominally normal hearing can have different requirements for broadcast speech to be both intelligible and comprehensible.

Initial clean audio work has focused on algorithmic enhancement of broadcast speech at the receiver end [11, 12]. When performed blindly with pre-mixed broadcast streams, these approaches have been shown to do little to improve speech intelligibility [13, 14], though some have demonstrated reduction in listener effort [15]. Recent approaches have exploited advancing audio technology and the associated standards e.g. the enhancement of the center channel of 5.1 audio, which is commonly used for speech only, to improve intelligibility for HoH listeners [16]. One study, utilizing new standards in spatial audio, showed an improvement in listener sentence recognition accuracy from 34% to 81% in the presence of 'applause' type noise [17].

It has been shown that an increase in speech volume of 3-6dB relative to the remaining audio offers improved intelligibility for HoH listeners [12]. However, most research has treated all remaining audio elements as maskers, irrespective of their salience or narrative importance to speech. Until recently, this assumption has not required challenging as standard linear broadcasting prevented separate control of non-speech audio objects. The advent of Object Based Audio (OBA) changes this. In OBA broadcasting, different audio elements are treated as separate 'objects' and are broadcast separately along with meta-data which is used to reconstruct the objects at the end-user [18, 19]. OBA technology presents a uniquely controllable broadcast environment which can not only provide users with a better broadcast experience but can, as this work will do, be used to facilitate investigation of the exact composition of optimally intelligible and comprehensible broadcast speech.

Literature distinguishes between two forms of speech intelligibility; signal-dependent (using 'bottom-up processing'), where the ability to retrieve the message is based solely on the speech signal, and complementary, which utilizes other non-speech cues from the speech signal, such as syntax and semantics as well as non-speech cues such as facial expressions [20]. These latter cues are also referred to as top-down information [21] and they have been shown to play a greater role in speech perception when hearing is challenged; either by impairment or masking from competing sources [22]. It is theorized that this is due, in part, to the manner in which the brain composes perceptual auditory objects, using the expected representation of the object to predict parts of the object for which no input is currently available [23]. This theory is consistent with results demonstrating the improvement in intelligibility which is achieved when the speech is linguistically predictable [21]. This is a widely replicated result, often quantified by Bilger's Revised Speech Perception in Noise tool (R-SPIN) [24, 25, 26] or variations thereof [3, 27].

Whilst the amount of linguistic context in broadcast speech cannot be controlled at point of service to it more intelligible, by utilizing OBA methods other audio elements which establish the context of this speech can be. This motivates research which aims to understand the effect different types of non-speech audio elements, like music, ambiances or foreground sound effects, have on intelligibility and comprehension of broadcast speech. Additionally it motivates understanding whether the strategic inclusions of these elements can be leveraged to improve intelligibility. Thus far, limited research has investigated this in a media environment; a 2000 study by Moreno showed for instructional messages additional audio elements can overload the listeners working memory [28]. Moreno's work however did not address what effects these elements have when the listener's auditory working memory is already under strain from impairment or from a complex acoustic scenario.

1.1. Research Questions

This doctoral work aims to answer two key research questions;

1. How do non-speech audio elements affect speech intelligibility and comprehension in broadcast material for HoH listeners?

2. How can this understanding be utilized intelligently, at point of service, to optimize intelligibility and comprehension of broadcast content for individual listeners?

2. Experimental Design

Theories of perceptual auditory object formation are gaining prominence as an explanation for speech perception abilities in *'cocktail party problem'* scenarios [29, 30]. These theories state that auditory objects are perceived by grouping acoustic features from the incoming auditory stream into stable spectro-temporal entities [23]. Given OBA's treatment of audio elements as objects, this technology presents a natural platform for exploring how the composition of audio objects (external to the listener) effects intelligibility of speech and comprehension of content by the listener.

This doctoral work builds on existing OBA research for HoH listeners from the University of Salford [31, 32] that shows, whilst a need for clearer speech is consistently present, the preferences of individual HoH listeners vary considerably with respect to non-speech elements. Specifically, this work will utilize a mixed-method paradigm to obtain a complete representation of listener's interaction with broadcast material; concurrently addressing objective measures of intelligibility and comprehension as well as subjective measures such as comfort and perceived quality.

This work hypothesizes that in the context of broadcast audio, the perceptual auditory objects which provide intelligible speech and comprehension of the content do not solely consist of speech but are also formed from non-speech elements which convey context and narrative meaning. It is intended that by determining the salient components which are used to form meaningful perceptual auditory objects when engaging with broadcast material, this parsing of separate audio elements into auditory objects can be outsourced to the OBA hardware itself. It is hoped that this will reduce listening effort for HoH listeners and, thus, improve overall understanding.

2.1. Current Work

The current experimental work aims to address research question one and constitutes two phases; the first investigates the effect salient non-speech elements have on intelligibility alone and the second will investigate both comprehension and intelligibility concurrently. This two phase approach is necessary as intelligibility is often assumed to be a valid proxy measure for comprehension, however the most recent and ecologically valid literature on the topic suggests that the correlation between these two measures is weak at best [33].

Phase one utilizes Bilger's R-SPIN tool [24] which has been modified to include salient non-speech audio elements, henceforth referred to as sound effects (SFX). These are introduced into half of the presented sentences and, for example, include a dog bark SFX for the item *my son has a dog for a pet*. The level of the competing babble noise from the test has been adjusted such that the combination babble plus SFX noise presents an equivalent energetic masker as the babble-only noise. This means energetic masking effects can be partialled out, allowing for determination of whether the additional non-speech elements behave as informational maskers or aid perceptual auditory object formation. If it is the latter, given the presence of sentences with both high and low linguistic predictability, comparison between the effects of different contextual cues (linguistic or non-speech audio) on intelligibility will be possible.

If these elements appear to have a masking effect, analysis will be undertaken to determine whether these additional elements are acting as bottom-up or top-down informational maskers. Bottom-up informational maskers are characterized by having similar spectral qualities to speech [34]. Spectral analysis of each SFX as compared to the spoken sentence and the speech band more broadly will be completed to determine whether there correlation between performance and elements with greater potential to be bottom-up maskers. If not, this will motivate further investigation of whether the added elements function as top-down informational maskers (presenting a cognitive distraction or overload [34]). Pupillometry will be utilized in this investigation to confirm and quantify the hypothesized cognitive load increase imposed by the elements acting as top-down informational maskers.

This work is ongoing and it is anticipated that collection and analysis of initial results, with normal hearing listeners, will be completed by mid-late August. These results and their analysis will inform whether any alterations to experimental design are required before collection of data from HoH listeners.

The second phase will utilize audio-visual content containing segments of OBA broadcast material which constitute complex auditory scenes. After viewing each segment, participants will be asked to repeat the final word or phrase in the dialogue (to quantify speech intelligibility, in the manner of R-SPIN style tests) and be asked to respond to a number of comprehension questions. A key focus for this stage will be ensuring that the stimulus is ecologically valid for a broadcast context. Most of the comparable research which has been done on the effects of non-speech elements has been in the context of multimedia learning [28], for normal hearing learners. Creation of this stimulus will be undertaken in conjunction with the University of Salford's School of Arts and Media and other media partners.

Correlation analysis of these results will aim to, first, elucidate whether there exists a sufficiently strong relationship between intelligibility and comprehension for the former to act as a proxy measure for both. Secondly, it will provide measures of the effects non-speech audio elements have on both comprehension and intelligibility in a broadcast context. This work hypothesizes that whilst the inclusion of additional audio elements may display some top-down informational masking effects, their inclusion will aid overall comprehension the content's narrative.

2.2. Future Work

To address the second research question, ongoing data collection will occur throughout the doctoral work to develop a database of HoH listener characteristics (which will be made open-access). This database will include objective data; pure tone audiometry thresholds, performance scores for R-SPIN and data about the individual listeners, such as age. It will also include subjective data about the listener's perceptions of optimally intelligible broadcast speech. The latter will be gathered qualitatively using self-report measures (surveys and semi-structured interviews). Complementary quantitative information about these subjective measures will be obtained in the manner of [32] where listeners are presented with broadcast material on an OBA platform which they can personalize to what they perceive as the optimal balance between speech, music, foreground and background sound effects for them.

Statistical analysis will be performed on these results to determine which individual characteristics of the listener has the strongest predictive power for their optimal intelligibility and comprehension requirements. This will form the basis of engineering work to develop an intelligent solution for acquiring this data from an end-user in order to automatically calibrate for their listening needs.

3. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] A. Bronkhorst and R. Plomp, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3132–3139, 1992.

[3] M. K. Pichora-Fuller, B. A. Schneider, and M. Daneman, "How young and old adults listen to and remember speech in noise," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 593–608, 1995.

[4] Action on Hearing Loss. (2015) Hearing Matters Report. [Accessed: 25/07/2016]. [Online]. Available: https://www.actiononhearingloss.org.uk/~/media/Documents/Policy%20research%20and%20influencing/Research/Hearing_Matters_2015/Hearing%20Matters%20Report.ashx

[5] H. Fuchs, S. Tuff, and C. Bustad, "Dialogue enhancement–technology and experiments," *EBU Technical Review*, vol. 2, pp. 1–11, 2012.

[6] Royal National Institute for Deaf People (RNID), "Annual survey report 2008," 2008.

[7] D. Cohen. (2011, March) Sound matters. BBC College of Production. [Accessed: 16/06/16]. [Online]. Available: http://www.bbc.co.uk/academy/production/article/art20130702112136134

[8] J. Paulus, J. Herre, A. Murtaza, L. Terentiv, H. Fuchs, S. Disch, and F. Ridderbusch, "Mpeg-d spatial audio object coding for dialogue enhancement (saoc-de)," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[9] Royal National Institute for Deaf People (RNID), "Annual survey report 2005," 2005.

[10] D. Ruggles and B. Shinn-Cunningham, "Spatial selective auditory attention in the presence of reverberant energy: individual differences in normal-hearing listeners," *Journal of the Association for Research in Otolaryngology*, vol. 12, no. 3, pp. 395–405, 2011.

[11] M. Armstrong, "Audio processing and speech intelligibility: a literature review," Tech. Rep., BBC, 2011.

[12] T. Komori, A. Imai, N. Seiyama, R. Takou, T. Takagi, and Y. Oikawa, "Development of volume balance adjustment device for voices and background sounds within programs for elderly people," in *Proc. Audio Engineering Society Convention 135*, New York, U.S.A., 2013.

[13] A. Carmichael, "Evaluating digital on-line background noise suppression: Clarifying television dialogue for older, hard-of-hearing viewers," *Neuropsychological rehabilitation*, vol. 14, no. 1-2, pp. 241–249, 2004.

[14] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

[15] H. Müsch, "Aging and sound perception: Desirable characteristics of entertainment audio for the elderly," in *Proc. Audio Engineering Society Convention 125*, San Francisco, U.S.A., 2008.

[16] B. Shirley and P. Kendrick, "The clean audio project: Digital tv as assistive technology," *Technology and Disability*, vol. 18, no. 1, pp. 31–41, 2006.

[17] H. Fuchs and D. Oetting, "Advanced clean audio solution: Dialogue enhancement," *Motion Imaging Journal, SMPTE*, vol. 123, no. 5, pp. 23–27, 2014.

[18] J. Popp, M. Neuendorf, H. Fuchs, C. Forster, and A. Heuberger, "Recent advances in broadcast audio coding," in *Proc. of 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, London, U.K., 2013, pp. 1–5.

[19] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilper, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer *et al.*, "Mpeg spatial audio object codingthe iso/mpeg standard for efficient coding of interactive audio scenes," *Journal of the Audio Engineering Society*, vol. 60, no. 9, pp. 655–673, 2012.

[20] N. Miller, "Measuring up to speech intelligibility," *International Journal of Language & Communication Disorders*, vol. 48, no. 6, pp. 601–612, 2013.

[21] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *The Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1337–1351, 1977.

[22] A. A. Zekveld, M. Rudner, I. S. Johnsrude, D. J. Heslenfeld, and J. Rönnberg, "Behavioral and fmri evidence that cognitive ability modulates the effect of semantic context on speech intelligibility," *Brain and language*, vol. 122, no. 2, pp. 103–113, 2012.

[23] J. K. Bizley and Y. E. Cohen, "The what, where and how of auditory-object perception," *Nature Reviews Neuroscience*, vol. 14, no. 10, pp. 693–707, 2013.

[24] R. Bilger, "Speech recognition test development," *ASHA Reports*, vol. 14, pp. 2–15, 1984.

[25] D. J. Schum and L. Matthews, "SPIN test performance of elderly hearing-impaired listeners," *Journal of the American Academy of Audiology*, vol. 3, no. 5, pp. 303–307, 1992.

[26] L. E. Humes, B. U. Watson, L. A. Christensen, C. G. Cokely, D. C. Halling, and L. Lee, "Factors associated with individual differences in clinical measures of speech recognition among the elderly," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 2, pp. 465–474, 1994.

[27] R. H. Wilson, R. McArdle, K. L. Watts, and S. L. Smith, "The revised speech perception in noise test (r-spin) in a multiple signal-to-noise ratio paradigm," *Journal of the American Academy of Audiology*, vol. 23, no. 8, pp. 590–605, 2012.

[28] R. Moreno and R. E. Mayer, "A coherence effect in multimedia learning: The case for minimizing irrelevant sounds in the design of multimedia instructional messages." *Journal of Educational Psychology*, vol. 92, no. 1, p. 117, 2000.

[29] K. T. Hill and L. M. Miller, "Auditory attentional control and selection during cocktail party listening," *Cerebral cortex*, 2009, [Accessed: 25/07/2016]. [Online]. Available: http://cercor.oxfordjournals.org/content/early/2009/07/02/cercor.bhp124.full

[30] S. Carlile, "Active listening: speech intelligibility in noisy environments," *Acoustics Australia*, vol. 42, no. 2, pp. 90–96, 2014.

[31] B. Shirley and R. Oldfield, "Clean audio for TV broadcast: An object-based approach for hearing-impaired viewers," *Journal of the Audio Engineering Society*, vol. 63, no. 4, pp. 245–256, 2015.

[32] B. Shirley, J. Woodcock, M. Meadows, and A. Tidball, "Personalized object-based audio for hearing impaired TV viewers," 2016, [Unpublished].

[33] L. Fontan, J. Tardieu, P. Gaillard, V. Woisard, and R. Ruiz, "Relationship between speech intelligibility and speech comprehension in babble noise," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 3, pp. 977–986, 2015.

[34] S. Carlile, "Auditory perception: Attentive solution to the cocktail party problem," *Current Biology*, vol. 25, no. 17, pp. 757–759, 2015.