INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

The effect of situation-specific non-speech acoustic cues on the intelligibility of speech in noise

Lauren Ward[1], *Ben Shirley*[1], *Yan Tang*[1], *William J. Davies*[1]

[1]Acoustics Research Centre, University of Salford, Manchester, U.K.

L.Ward7@edu.salford.ac.uk, {B.G.Shirley, Y.Tang, W.Davies}@salford.ac.uk

Abstract

In everyday life, speech is often accompanied by a situation-specific acoustic cue; a hungry bark as you ask *'Has anyone fed the dog?'*. This paper investigates the effect such cues have on speech intelligibility in noise and evaluates their interaction with the established effect of situation-specific semantic cues. This work is motivated by the introduction of new object-based broadcast formats, which have the potential to optimise intelligibility by controlling the level of individual broadcast audio elements, at point of service. Results of this study show that situation-specific acoustic cues alone can improve word recognition in multi-talker babble by 69.5%, a similar amount to semantic cues. The combination of both semantic and acoustic cues provide further improvement of 106.0% compared with no cues, and 18.7% compared with semantic cues only. Interestingly, whilst increasing subjective intelligibility of the target word, the presence of acoustic cues degraded the objective intelligibility of the speech-based semantic cues by 47.0% (equivalent to reducing the speech level by 4.5 dB). This paper discusses the interactions between the two types of cues and the implications that these results have for assessing and improving the intelligibility of broadcast speech.

Index Terms: intelligibility, speech perception, broadcast media

1. Introduction

Situational awareness shapes much of an individual's perception of speech, particularly in noisy or challenging listening scenarios. Situation-specific cues from within the speech itself, such as semantic context, as well as non-speech cues aid in decoding the intended message [1, 2]. This occurs largely by facilitating the listener's predictions of segments of speech which have been obscured by noise or for which no signal is available [3]. Whilst the effect of semantic contextual cues are well understood [4, 5, 6, 7], recent work in the area has highlighted the need for better characterisation of the complex relationship between different types of contextual cues [1]. Understanding of the effects of situation-specific acoustic cues, in particular their interaction with other cues, is limited [8, 9]. This paper endeavours to address this need, investigating the effects that situation-specific acoustic cues, and their interaction with semantic cues, have on speech intelligibility in noise.

Situation-specific acoustic cues figure prominently in broadcast media. Such cues, in the form of sound effects (SFX), Foley or ambiances, are routinely included to create realism as well as themselves being narratively important and progressing the plot [10]. Research to date on broadcast speech intelligibility has taken a binary view of the audio; speech versus competing noise. Subsequently methods to assess or improve intelligibility have focused on speech enhancement [11]. Recent developments in broadcast technology, specifically the introduction of object-based audio, has meant that these types of acoustic cues can be treated as independent 'audio objects'. As separate objects, how they are reproduced by the end-user's TV receiver at point-of-service can be altered based on their own metadata and is independent of other audio objects [12]. This metadata facilitates an increasing amount of knowledge about the characteristics of these objects including their location and, potentially, their narrative importance [13]. Thus, the question of what effect these situation-specific acoustic cues have on intelligibility, and how they should be reproduced at point of service for optimal intelligibility, is a pertinent one [10].

The effect of semantic contextual cues on speech intelligibility in noise has been widely replicated [5, 14, 15, 16]. This effect is commonly quantified using the Revised Speech Perception in Noise (R-SPIN) test [4, 17, 18] or variations thereof [1, 5, 19, 20]. R-SPIN utilises sentences with controlled predictability; containing either low or high amounts of semantic context. R-SPIN results have shown that word recognition is, on average, 40-45$\%$ higher when semantic context is present [20]. Recent work has utilised R-SPIN to demonstrate the interaction between semantic cues and pictorial situation-based cues on listening and lip-reading [1]. Both types of contextual cues have proven to be beneficial, but with different influences on speech perception. A number of other studies have investigated the interaction between semantic cues and visual cues [2, 21]. However, there has been little research investigating the effect of non-speech acoustic cues on intelligibility. The most recent, related study showed that for urgent, public-address style speech, preceding sounds cues can positively influence the intelligibility [8]. Similar concepts have been explored in studies of knowledge transfer in multimedia learning yielding different results; a study by Moreno in 2000 showed that, for instructional messages, additional audio elements can overload the listeners' working memory [9].

This work investigates the effect of situation-specific acoustic cues on speech intelligibility in multi-talker babble. Section 2 outlines the experimental design, with the perceptual results in Section 3 ([22] reports work-in-progress results from a smaller cohort). Assessing the inclusion of acoustic cues presents an additional challenge, compared with semantic cues, as the acoustic cues themselves have the potential to act as maskers. Objective intelligibility measures can facilitate analysis how signal and masker interactions affect word level intelligibility (e.g. [23, 24]). One such measure, the glimpse proportion (GP) [25], quantifies the number of time-frequency regions of speech which survive energetic masking and is intended to reflect the local audibility of speech in noise. To this end, section 4 undertakes an objective intelligibility analysis using the GP measure to evaluate the potential signal-level interaction between acoustic and semantic cues. Section 5 discusses the implications of the results for intelligible broadcast content whilst Section 6 draws conclusions and outlines future work.

http://dx.doi.org/10.21437/Interspeech.2017-500

2. Experimental Design

To evaluate the effects of situation-specific acoustic cues, with reference to the known effect of semantic cues, a modified version of the R-SPIN test was employed (development and validation of the original R-SPIN test is in [4, 14]). This approach also facilitated evaluation of any interaction between the cues.

2.1. Stimuli

The R-SPIN stimuli consist of short, phonetically balanced sentences spoken by a male speaker in American English and presented in multi-talker babble. All sentences end with a monosyllabic noun, the keyword, which participants are scored on their ability to correctly identify. The original test evaluates a single factor, the effect of priming the listener with situation-specific semantic cues (conditions LP and HP in Table 1). This is achieved by controlling the predictability of the sentences, either giving the listeners clues to the keyword through high predictability (HP) sentences (e.g. *'Stir your coffee with a **spoon**'*) or no clues through low predictability (LP) sentences (e.g. *'Bob could have known about the **spoon**'*). The LP sentences are the control condition in this experiment. This was maintained from the original test, as opposed to presenting the keyword in isolation, as the preceding sentences affords the participant the same opportunity to direct their attention towards the target stimuli as in the other conditions. This work's modified version adds a second factor: acoustic cues conveyed by SFX (conditions LP+SFX and HP+SFX seen in Table 1). These SFX were introduced into half of the presented sentences. All SFX ended prior to the keyword being spoken (as seen in Figure 1), to ensure both types of situation specific cues had equal opportunity to prime the listener.

Table 1: *Four conditions of the modified R-SPIN*

Condition	Factor 1	Factor 2	# Stimuli
LP	Low Predictability	No SFX	50
HP	High Predictability	No SFX	50
LP+SFX	Low Predictability	SFX	50
HP+SFX	High Predictability	SFX	50

The R-SPIN materials were taken from the original CD recording. Any CD artefacts were cleaned using Adobe Audition. The multi-talker babble from the original stimuli, mixed from 12 speakers, was maintained as the masker to allow direct comparison of the results of this study with previous studies. Furthermore, multi-talker babble was considered suitably representative of the type of masker which a listener may encounter in broadcast content, as maskers in broadcast content often fluctuating and overlap the speech spectrum. The original R-SPIN corpus contains 400 sentences organised into 8 lists of 50. Four of these lists were selected; lists 1, 2, 5, and 6, constituting 200 sentences. In these 200 sentences, each keyword is presented twice, once with semantic cues and once without giving 100 unique keywords. 50 of these keywords are shared between conditions LP and HP and 50 are shared between conditions LP+SFX and HP+SFX. Usage of different keywords in each condition pair was designed to reduce learning effects. Some sentences were swapped with those from other lists to ensure half the keywords could be paired with suitable SFX, however as much of the original list integrity as possible was maintained.

The SFX selected were taken from broadcast quality SFX libraries (BBC Sound Effects Library [26] and Soundsnap [27]).

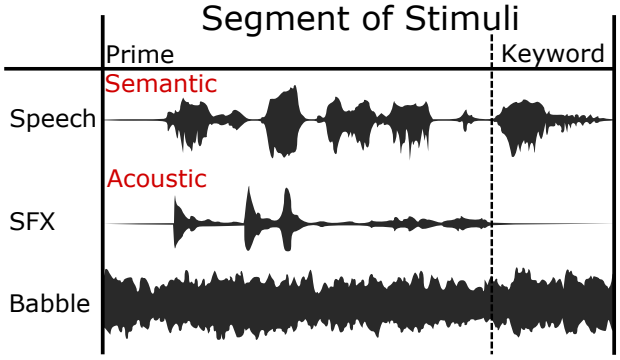

Figure 1: *Example stimuli, noting alignment of the priming cues*

They were chosen to have equivalent priming effects as the semantic cues in the HP sentences. For example, the HP sentence for the keyword *pet* is *'My son has a dog for a pet'*, which utilises the assumed knowledge that children often keep pet dogs. As such, the SFX selected for this keyword was a dog's bark. The SFX used were not limited to recordings of the keyword itself but also used combinations of sounds e.g. the SFX for the keyword *pond* consisted of the sounds of water splashing and ducks quacking.

2.2. Implementation

Given this research's broadcast context signal loudness was measured as it would be in the production of broadcast content, using the ITU-R BS.1770-2 algorithm for measuring broadcast programme loudness, noted as dB^{LKFS} [28]. This algorithm utilises a K-weighted filter before calculating the mean square of the signal over the signal duration (excluding periods of quiet). -23 dB^{LKFS} was selected as the target level, as this is the standard level for broadcast audio in the U.K. [29]. The speech and multi-talker babble from the original R-SPIN materials were each normalised to this level. The SFX were also normalised to -23 dB^{LKFS}. The SFX were then combined with the babble only signal, for conditions LP+SFX and HP+SFX, and this mixture was normalised to -23 dB^{LKFS}, giving the masking signal in all conditions equivalent energetic masking potential.

This experiment maintained the original test's single signal to babble ratio (SBR) paradigm [4]. A static SBR with a single speaker and consistent speaking pace throughout reduced rendered the different situation-specific cues as the only salient differences between the stimuli. To set an appropriate SBR, which would yield a speech reception threshold of approximately 50% for conditions from the original test (LP and HP) a small pilot study was undertaken with experienced listeners ($n = 4$). From this an SBR of -2 dB was selected, providing an average word recognition rate (WRR) of 53.5% across conditions LP and HP. The pilot study was also used to verify that any effects caused by acoustic cues on WRR in conditions LP+SFX and HP+SFX did not result in floor or ceiling effects [14]. Results showed a WRR of 72.5% and 80.0% for condition LP+SFX and HP+SFX respectively, avoiding ceiling effects.

The sentences and babble+SFX mixture were co-located and presented from a Genelec 8030A Studio Monitor mounted at a height of $1.1m$. The listener was situated $2.2m$ from the speaker and encouraged sit relatively still, as if viewing broadcast content. The study was undertaken in a listening room meeting the ITU-R BS.1116-1 standard for listening tests

[30]. The stimuli were presented at a sound pressure level of 69 dB(A), measured at the listening position (as in similar studies [18, 31]). Participants were presented with 15 practice stimuli to allow them to develop familiarity with the task and the speaker's voice. All participants were presented with the stimuli in pseudo-randomised order, with the presentation order counterbalanced across the participant pool to avoid learning effects.

2.3. Study Participants

24 native English speakers (7 females and 17 males) who had self-reported normal hearing participated (not inclusive of pilot study participants). 54% of participants were aged $18-29$ yrs, 29% aged $30-39$ yrs, 17% aged 40 or older. They had a variety of listening test experience: half of the participants were naive listeners, 21% with moderate experience and 29% were experienced participants of listening tests.

3. Perceptual Intelligibility Results

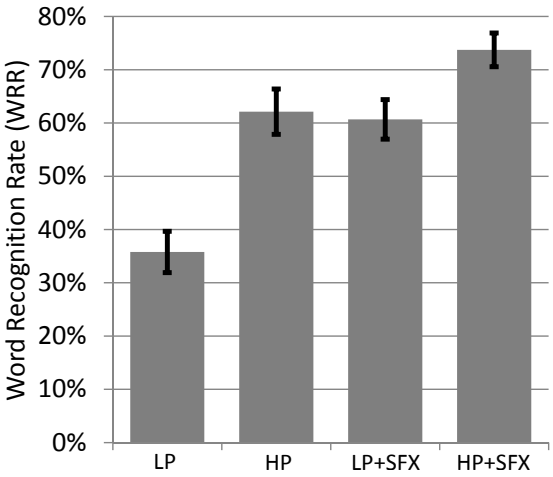

Figure 2: *Mean word recognition rate ($n = 24$) for each experimental condition. Error bars indicate ± 1 standard error.*

Figure 2 shows the mean WRR with standard error bars. Condition LP has the lowest mean WRR of 35.8%. Condition HP gives a 73.5% improvement in recognition relative to condition LP, increasing the WRR to 62.1%. This result is consistent with other implementations of the R-SPIN test [1, 20]. For condition LP+SFX the WRR increases to 60.7%. This improvement of 69.5%, given it's similarity in magnitude to the improvement in condition HP, suggests that acoustic cues offer similar levels of benefit to intelligibility as semantic cues. Condition HP+SFX shows a WRR of 73.7%, an improvement of 106.0% from condition LP and a 21.5% and 18.7% improvement from the conditions with only acoustic and semantic cues respectively. This indicates that the combined effect of the cues also yields a modest improvement in intelligibility compared with either cue in isolation.

To determine the significance of these effects, a two-way repeated measures ANOVA was performed. The effects of both types of cues were significant with $[F = 240.53, p < 0.001]$ and $[F = 127.34, p < 0.001]$ for semantic and acoustic cues respectively. Of particular note is that the interaction between the semantic and acoustic cues was also statistically significant $[F = 21.32, p < 0.001]$. Whilst the effect between the participants was also weakly statistically significant $[F = 11.28, p < 0.005]$, it is likely not practically significant given its small F-ratio compared with the major effects. Post-hoc pairwise comparisons of the significant conditions were performed using Tukey's Honest Significant Difference test. This showed that all pairs of conditions were significantly different from each other [all $p < 0.001$], with the exception of the pair of conditions HP and LP+SFX [$p = 0.17$].

4. Objective Intelligibility Analysis

Section 3 indicates that, in the absence of other cues, the presence of acoustic cues has a similar benefit as semantic cues for speech intelligibility. In order to investigate the contributions from the semantic and acoustic cues, the objective intelligibility was measured using the GP [25]. The GP was independently calculated for the priming speech and the keyword in each stimulus. The average scores across the 50 sentences in each condition are presented in Table 2. For the keyword, the objective intelligibility measured as GP is very similar in the four conditions, and statistical analysis suggested no significant differences between any of the conditions [all $p > 0.05$]. From this it appears that energetic masking on the keyword was, on average, equivalent across the conditions. This is consistent with differences in WRR depending on different priming cues.

Table 2: *Glimpse proportion (GP) for each condition's keyword and priming speech. Parentheses indicate standard error.*

	LP	**HP**	**LP+SFX**	**HP+SFX**
Keyword	13.56% (0.76)	12.53% (0.82)	14.23% (0.84)	12.42% (0.83)
Priming Speech	18.86% (0.61)	18.57% (0.53)	9.85% (0.44)	10.07% (0.53)

The objective intelligibility measures used here are signal-driven, meaning they are representative of factors which affect the speech intelligibility at the signal level such as signal to noise ratio and envelope modulation of speech. For this reason the GP was also calculated over the priming speech in the stimuli in order to facilitate analysis of the signal level effect of the acoustic cues. From Table 2 it can be seen that for the priming speech, the GP in the conditions with semantic cues only (LP and HP) are very similar, as well as in the conditions with acoustic cues (LP+SFX and HP+SFX). This is in contrast to the large intelligibility gains which are observed from the changes in subjective WRR between conditions LP and HP, and LP+SFX and HP+SFX (Figure 2). Furthermore, the GP in the latter cases is significantly lower, with a reduction of 47.0% on average $[F = 272.44, p < 0.001]$. This degradation in objective intelligibility was corroborated using a standard intelligibility measure, the Speech Intelligibility Index (SII) [32], which showed average $SII = 0.36$ for conditions with semantic cues only and $SII = 0.24$ for those with acoustic cues, a degradation of 31.6% (statistically significant with $[F = 297.04, p < 0.001]$).

Having observed the large improvements that each cue separately imparted (HP and LP+SFX in Figure 2) the combined improvement, condition HP+SFX, is far more modest in comparison. This suggests that one of the cues may be impaired by the existence of the other, assuming that the benefits of the semantic and acoustic cues are additive and ceiling effects were not reached. One possibility is that the introduction of the acoustic cue leads to the reduction of the effective SBR, hence degrading intelligibility of the priming speech. Consequently,

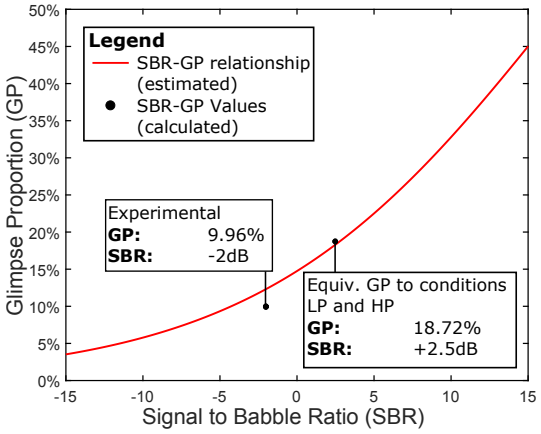

Figure 3: *Equivalent SBR for condition HP+SFX required to achieved $GP = 18.72\%$ for the priming speech, shown with relationship between GP and SBR*

the semantic cues available to the listener are compromised by the acoustic cues at the signal level, resulting in reduced benefit. The amount that the speech would need to be increased by to compensate for this effective SBR reduction was investigated. By increasing the level of the priming speech and calculating the respective GP, the SBR required to meet $GP = 18.72\%$ (as in conditions LP and HP) was found to be $SBR = +2.5$ dB. This is shown in Figure 3 along with the GP for the experimental $SBR = -2$ dB. A 4.5 dB increase from the experimental SBR is observed. To determine whether the 4.5 dB increase is likely to translate to other SBR, the relationship between SBR and GP is also plotted in the range -15 dB to $+15$ dB (found by fitting a two-parameter logistic function). Figure 3 indicates that a 4.5 dB increase is likely to compensate for effective SBR reduction at higher SBR. Below the experimental $SBR = -2$ dB the floor conditions are quickly reached, indicating the 4.5 dB value may not hold in this region.

5. Further Implications

These results indicate that the relationship between dialogue, background noise and sound effects in broadcast content is significantly more complex than previously addressed. As most broadcast speech is presented with some semantic context, particularly for narrative based programs where SFX are most commonly used, the results from condition HP+SFX have the greatest implications for broadcast media. The modest, but significant 18.7% increase in intelligibility offered by the acoustic cues when semantic cues are already present must be considered in the context of possible signal degradation effects. Furthermore, it is possible that there is also competition between the two cues for the listener's attention when being presented simultaneously. In the presence of the acoustic cues, the listener's attention to the semantic cues in the priming speech may be distracted, negatively affecting the allocation of cognitive resources to process the semantic cues. This may result in the listener switching their attention between the different cues over the time, in order to process the information from both cues, which may increase cognitive load of the listener [33]. This increased load potentially impairs the parsing of each cue, compared with when the cognitive resources are mostly dedicated to one cue at a time. This is a similar hypothesis as was proposed in [9], where the addition of music and SFX was shown to reduce knowledge transfer in multimedia content.

In most broadcast content further situation-specific cues are present through the visual modality and spatial information. Whilst these additional cues may reduce the significance of any signal-level degradation presented by acoustic cues, they may further increase the potential for cognitive attention to be negatively impacted. There are caveats on the generalisability of the current results, given that broadcast content is rarely reproduced monaurally and multi-talker babble is only one specific masker type. Despite this, a number of specific applications exist where consideration of the magnitude of this dual effect of acoustic cues is particularly pertinent. These include radio dramas and the provision of audio description for visually impaired TV viewers, in which the visual modality is compensated for with greater amounts of speech and, potentially, greater numbers of acoustic cues [34] which may overlap this speech.

The key result here is that if intelligibility is to be accurately assessed or improved in broadcast content, the influence of situation-specific acoustic cues and their interaction with other contextual cues cannot be ignored. However, their consideration requires knowledge of their presence as well as their narrative importance and saliency. Object-based audio technology offers a way to integrate information about discrete audio events into intelligent assessments of the intelligibility of broadcast speech. However, gaining and utilising such knowledge in intelligibility fields outside of broadcast media still presents a real challenge. For these fields, the work presented here contributes fundamental understanding, which may help to to explain intelligibility results obtained in scenarios where salient situation-specific acoustic cues are present.

6. Conclusion

The presence of situation-specific acoustic cues, when no other contextual cues are present, was shown to increase word recognition in noise by 69.5%. The combination of both semantic and acoustic cues further improved performance by 106.0%, compared with no cues and 18.7%, compared with semantic cues only. However, the presence of acoustic cues degraded glimpsing opportunities by up to 47.0%. To compensate for this, speech would need to be increased by 4.5 dB relative to the masker. These results have significant implications for understanding the perception of speech in scenarios where situation-specific cues may be present as well as for the creation of intelligible broadcast speech.

A significant amount of work is still required in order to evaluate the possible signal-level and cognitive attention effects of situation-specific acoustic cues. Furthermore, the effects of cue saliency and their effects at different SBR and in the presence of different masker types are yet to be established. Ongoing work by this group aims to establish how the usage of acoustic cues may affect different listeners, in particular whether the results demonstrated here are replicated in hard of hearing populations. Future work will investigate how the cognitive and signal-level effects of acoustic cues alter at different masker/cue ratios, in order to develop more intelligent and personalisable assessments of broadcast intelligibility.

7. Acknowledgments

Lauren Ward is supported by the General Sir John Monash Foundation. The authors would also like to acknowledge the contributions of Huw Swanborough and Philippa Demonte.

8. References

[1] B. Spehar, S. Goebel, and N. Tye-Murray, "Effects of context type on lipreading and listening performance and implications for sentence processing," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 3, pp. 1093–1102, 2015.

[2] M. Aguert, V. Laval, L. Le Bigot, and J. Bernicot, "Understanding expressive speech acts: the role of prosody and situational context in french-speaking 5-to 9-year-olds," *Journal of speech, language, and hearing research*, vol. 53, no. 6, pp. 1629–1641, 2010.

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

[4] R. Bilger, *Speech recognition test development, In: E. Elkins ed. Speech recognition by the hearing impaired*, 1984, vol. 14, pp. 2–15.

[5] S. Sheldon, M. K. Pichora-Fuller, and B. A. Schneider, "Priming and sentence context support listening to noise-vocoded speech by younger and older adults," *The Journal of the Acoustical Society of America*, vol. 123, no. 1, pp. 489–499, 2008.

[6] N. Golestani, S. Rosen, and S. K. Scott, "Native-language benefit for understanding speech-in-noise: The contribution of semantics," *Bilingualism: Language and Cognition*, vol. 12, no. 03, pp. 385–392, 2009.

[7] R. F. Holt and T. Bent, "Children's use of semantic context in perception of foreign-accented speech," *Journal of Speech, Language, and Hearing Research*, vol. 60, pp. 223–230, 2017.

[8] N. Hodoshima, "Effects of urgent speech and preceding sounds on speech intelligibility in noisy and reverberant environments," in *Proc. Interspeech 2016: 17th Annual Conf. of International Speech Communication Association*. San Francisco, U.S.A.: ISCA, 2016, pp. 1696–1699.

[9] R. Moreno and R. 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.

[10] M. Armstrong. (Oct, 2016) BBC white paper WHP 324: From clean audio to object based broadcasting. BBC. [Online]. Available: http://www.bbc.co.uk/rd/publications/whitepaper324

[11] ——. (Jan, 2011) BBC white paper WHP 190: Audio processing and speech intelligibility: a literature review. BBC. [Online]. Available: http://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP190.pdf

[12] J. Popp, M. Neuendorf, H. Fuchs, C. Forster, and A. Heuberger, "Recent advances in broadcast audio coding," in *Proc. 9^{th} IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*. London: IEEE, 2013, pp. 1–5.

[13] B. Shirley, M. Meadows, F. Malak, J. Woodcock, and A. Tidball, "Personalized object-based audio for hearing impaired TV viewers," *Journal of the Audio Engineering Society*, In Press.

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

[15] J. Aydelott and E. Bates, "Effects of acoustic distortion and semantic context on lexical access," *Language and Cognitive Processes*, vol. 19, no. 1, pp. 29–56, 2004.

[16] M. H. Davis, M. A. Ford, F. Kherif, and I. S. Johnsrude, "Does semantic context benefit speech understanding through top–down processes? evidence from time-resolved sparse fmri," *Journal of Cognitive Neuroscience*, vol. 23, no. 12, pp. 3914–3932, 2011.

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

[18] L. Humes, B. Watson, L. Christensen, C. Cokely, D. 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.

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

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

[21] A. A. Zekveld, M. Rudner, I. S. Johnsrude, J. M. Festen, J. H. Van Beek, and J. Rönnberg, "The influence of semantically related and unrelated text cues on the intelligibility of sentences in noise," *Ear and hearing*, vol. 32, no. 6, pp. 16–25, 2011.

[22] L. Ward, B. Shirley, and W. J. Davies, "Turning up the background noise; the effects of salient non-speech audio elements on dialogue intelligibility in complex acoustic scenes," in *Proc. of Institute of Acoustics 32nd Reproduced Sound Conf.* Southampton: IOA, Nov. 2016.

[23] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication*, vol. 49, no. 5, pp. 402–417, 2007.

[24] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Proc. Interspeech 2011: 12th Annual Conf. of International Speech Communication Association*. Florence, Italy: ISCA, 2011, pp. 345–348.

[25] M. Cooke, "A glimpsing model of speech perception in noise," *Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[26] BBC, "BBC Sound Effects Library CDs 1-60."

[27] Tera Media and CRG. Soundsnap.com. [Online]. Available: http://www.soundsnap.com/

[28] ITU Recommendation, "ITU-R BS. 1770-2, Algorithms to measure audio programme loudness and true-peak audio level," 2011.

[29] EBU, "Tech 3343 Guidelines for production of programmes in accordance with EBU R 128," Jan. 2016.

[30] ITU Recommendation, "ITU-R BS.1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

[31] B. Shirley, "Improving television sound for people with hearing impairments," Ph.D. dissertation, University of Salford, 2013.

[32] ANSI S3.5, "ANSI S3.5-1997 Methods for the calculation of the Speech Intelligibility Index," 1997.

[33] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention—focusing the searchlight on sound," *Current Opinion in Neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.

[34] M. Lopez, "Perceptual evaluation of an audio film for visually impaired audiences," in *Proc. 138th Audio Engineering Society Convention*. Warsaw, Poland: AES, 2015.