Audio Engineering Society

Convention Paper 10011

Presented at the 144th Convention
2018 May 23–26, Milan, Italy



# Speech-To-Screen: Spatial separation of dialogue from noise towards improved speech intelligibility for the small screen

Philippa J. Demonte[1], Yan Tang[1], Richard J. Hughes[1], Trevor J. Cox[1], Bruno M. Fazenda[1], Ben G. Shirley[1]

[1] *Acoustics Research Centre, University of Salford, Salford, M5 4WT, United Kingdom*

Correspondence should be addressed to Philippa Demonte (p.demonte@edu.salford.ac.uk)

**ABSTRACT**

Can externalizing dialogue when in the presence of stereo background noise improve speech intelligibility? This has been investigated for audio over headphones using head-tracking in order to explore potential future developments for small-screen devices. A quantitative listening experiment tasked participants with identifying target words in spoken sentences played in the presence of background noise via headphones. 16 different combinations of 3 independent variables were tested: speech and noise locations (internalized/externalized), video (on/off), and masking noise (stationary/fluctuating noise). The results revealed that the best improvements to speech intelligibility were generated by both the video-on condition and externalizing speech at the screen whilst retaining masking noise in the stereo mix.

## 1 Introduction

Small-screen devices, such as mobile phones and tablets, are increasingly being used to access audio-visual content [1][2]. However, the speech intelligibility of this content may be compromised by the energetic masking effects of a noisy listening environment or background noise on the audio soundtrack itself [3][4]. Furthermore, the use of headphones with these small-screen devices causes internalization effects, whereby sounds are incorrectly perceived to be emanating from inside the head instead of from an external source [5].

Our hypothesis is that by using binaural processing to place the dialogue track on the screen (externalized), i.e. co-located with the perceived location of the speaker, whilst maintaining other sounds in the stereo mix (internalized), creates spatial separation and a release of speech from masking, and hence improved intelligibility. Consequently, this work considers the case where both the dialogue and interfering noise come from the soundtrack.

It is well-established that separating speech and dialogue in azimuth can improve intelligibility [6][7]. The effect of separating speech and noise in terms of distance has been much less studied. Westermann and Buchholz (2013) [8] conducted an experiment into spatial separation and speech intelligibility with all audio to the front and distances of 0.5 m and 10.0 m from the head position. They found that noise binaurally auralized further away from speech resulted in higher intelligibility. They did not explore what happens when either the dialogue or noise are internalized, however.

Plail and Fazenda (2013) [9] conducted a study to quantify the perception of externalization. Competing speech signals were binaurally rendered at various externalized positions, with the internalized position as a control. They found that speech intelligibility significantly increased when the sources were at separate distances of 1.0 m and

1.5 m from the head position, but only in the lateral plane (±70° azimuth position), not in the frontal plane (±10° azimuth position). However, by their own admission, some subjects may not have externalized the sounds.

Recent industrial developments mean that it is becoming easier to render different sounds spatially in a mix. International broadcasters and film companies will increasingly be using object-based audio for improved accessibility and personalisation [10][11]. Swedish Radio, for example, has piloted a mobile phone app with 3.0 audio: speech was separated onto the centre channel whilst retaining the other audio in the regular stereo mix [12][13]. This centre channel was not binaurally externalized during the trial period, however. With the increasing availability of head-tracking systems, in the future binaural processing to increase the chances of externalization with headphones will become more common.

This paper presents focussed experiments to test whether separating the speech and background sounds, one being external, the other internal, can improve speech intelligibility. Section 2 outlines the methodology of a psychoacoustics listening experiment. Section 3 presents the results and statistical analysis. Plausible reasons for the results and further discussion regarding the applications of this research are outlined in Section 4.

## 2 Methodology

### 2.1 Experiment Overview

The experiment followed the widely-used method for testing speech intelligibility, where participants were required to identify target words in the presence of background noise [14][15][16]. Correct word scores were calculated for the collected data which were then statistically analyzed.

The experiment was conducted in a room acoustically treated to standard ITU-R BS.1116-1 [17]. Participants sat in the position within the room at which the Salford-BBC spatially-sampled Binaural Room Impulse responses (SBBCss BRIRs) [18] had been recorded at, and listened via STAX SR-2017 headphones to binaurally auralized [19] spoken sentences energetically masked by noise. During half of the experiment the corresponding video footage of the speakers was additionally presented on a monitor screen located directly in front of participants. The video clips were formatted to a 0.4 x 0.4 sized window, analogous to the size of a small-screen device, within a Matlab-generated Graphical User Interface (GUI). Outside of this window, the GUI also featured the instructions for the experiment and a virtual keyboard, which allowed participants a clear view when inputting responses. In order to be representative of small screen device viewing and to maximize the plausibility of the externalization effect, the monitor screen was placed at a distance of 1.0 m from the participants, matching the loudspeaker-to-microphone distance that the relevant BRIRs had been recorded at.

### 2.2 Speech Stimuli: GRID corpus

GRID [20] was chosen as the target speech corpus for this listening experiment, as it features 1000 audio-visual recordings by each of 18 male and 16 female British-English speakers. Of these, 8 male and 8 female speakers were selected by a process of elimination based on informal judgement of the criteria of: clarity of voice; consistent tempo and tone fall/rise of utterances; consistent head and shoulder framing on the videos, with the speakers facing straight towards the camera.

GRID corpus sentences comprise of a 6-word format: a verb, a colour, a preposition, a letter, a number, and a temporal word, for example: *"Place red in G4 now."* The grid references (letter-number combinations) were used as the target words for this experiment, as these occur in the middle of each GRID sentence and cannot be predicted when either energetically masked or not attended to. All the letters of the alphabet, except for 'w', and all the numbers from zero to nine feature in the corpus.

A total of 320 sentences, 20 sentences per speaker, were selected from the audio-visual recordings within the corpus, ensuring an even distribution of the letter-number combinations.

### 2.3 Variables

A total of 16 different combinations of 3 independent variables were tested in this listening experiment:

#### 2.3.1 Four binaural auralization positions:

- **INT**: both target speech and masking noise internalized at the headphones
- **SN**: target speech internalized at the headphones; masking noise externalized at the screen
- **NS**: masking noise internalized at the headphones; target speech externalized at the screen
- **EXT**: both target speech and masking noise externalized at the screen

For the internalization effect with headphones, uncorrelated signals were sent to the left and right ear, i.e. stereo reproduction was used. In contrast, to promote externalization of signals to the monitor screen whilst wearing headphones, the relevant SBBCss BRIR for the 0 degree position was used with the speech signal, whilst the relevant ±30 degree BRIRs were used with the masking noise. This was in order to realistically reproduce speech as a point source and masking noise as a diffuse source.

Participants were encouraged to remain facing forwards for the duration of the listening experiment. However, a head-tracking system was utilized, so that externalization of signals at the screen was more likely. Head movements were detected and monitored with a ceiling-mounted OmniTrack Trio system and spatial markers attached to the headphones. Based on these data, the BBC renderer [21] in real time convolved the audio with the relevant SBBCss BRIRs for the forward positions. Head-tracking calibration for the 0 degree position was conducted with participants at the start of the experiment.

#### 2.3.2 Two audio-visual playback conditions:

- **Audio-only**: video off
- **Video and Audio**: video on

Following acquisition of the audio-visual materials from the GRID corpus creators, it was ascertained that the audio on the video files had been captured via the in-built microphone on the camera, whilst the audio-only files had been simultaneously captured via a separate microphone. A cross-correlation and temporal shifting procedure therefore was applied in order to properly synchronize the audio with the corresponding videos for this listening experiment.

#### 2.3.3 Two types of noise maskers:

- **Speech-Shaped Noise (SSN)**: signal-to-noise ratio (SNR) set at -9 dB
- **Speech-Modulated Noise (SMN)**: SNR set at -12 dB

Speech signals were separately presented in two types of maskers: speech-shaped noise (SSN) and speech-modulated noise (SMN), representing temporally-stationary and temporally-fluctuating maskers respectively. To generate SSN, white noise is filtered using the coefficients of the long-term spectral envelope of the speech corpus, which are estimated by $10^{th}$-order linear predictive coding. Consequently, the long-term average spectrum (LTAS) of SSN matches that of the corpus. SMN is produced by applying the envelope extracted from a speech signal to the SSN signal in the time domain. This leads to the large temporal modulation of SMN; the spectrum of SMN, however, remains the same as for SSN. Figure 1 shows an example of the waveform for each masker accompanied by their LTAS.

The target speech-to-noise ratio (SNR) for each masker was chosen empirically to result in an intelligibility of between 40-60% when both the speech and masker were internalized, i.e. the baseline condition. This was to avoid the ceiling and flooring effects when the baseline performance was too high or too low. In order to offset the greater energetic masking effect of SSN versus increased opportunity for glimpsing target speech within SMN, a greater negative SNR was therefore required for target speech in SMN.

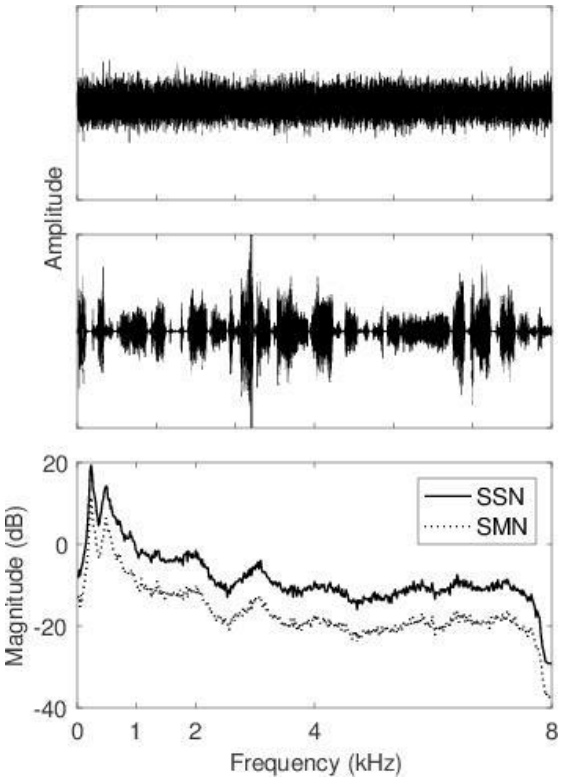


Figure 1. Sample waveforms of the maskers and their long-term average spectra. For illustration, the spectra of SSN and SMN are offset at ±3 dB respectively.

During the test, the intensity of the speech signals was normalized to the same root-mean square value. The presentation level for the target speech was calibrated and fixed at approximately 69 dB A using a B & K Type 2610 measuring amplifier and artificial ear. This chosen level falls within the normal range for conversation in quiet conditions. The relative levels of the straight-to-headphones-, externalized speech-, and externalized masker feeds were then calibrated to the same dB A using pink noise. The head-tracking system was turned off during this time in order to obtain the binaural calibration values for the front-facing BRIRs.

### 2.4 Experiment Structure

The experiment comprised of 4 sessions: audio-only and SSN masking noise; audio-only and SMN masking noise; video-on and SSN masking noise; video-on and SMN masking noise. The playback order of the 4 sessions, the 4 binaural auralization positions within each session, the speakers, and the sentences were randomized using a GUI designed in Matlab, which also captured the data entered by the participants via a keyboard.

| | | **MASKERS** | | | |
|---|---|---|---|---|---|
| | | SSN | | SMN | |
| **PRESENTATION** | Audio only | INT | SN | INT | SN |
| | | NS | EXT | NS | EXT |
| | Video + audio | INT | SN | INT | SN |
| | | NS | EXT | NS | EXT |

Table 1. Overview of the 16 combinations of 3 independent variables tested: binaural auralization positions; video on/off; masking noises.

Prior to the start of the experiment, participants were provided with a practice session comprising of 10 additional GRID sentences with examples of the playback conditions. During the main experiment a total of 320 speech-in-noise sentences were played once only. The data entered by each participant comprised of 20 pairs of letter-number grid references for each of the 16 combinations of the 3 independent variables.

### 2.5 Participants

20 native British-English speakers between the ages of 18-35 with self-reported normal hearing were recruited and paid for their participation in the listening experiment.

## 3 Results

Using a Matlab script, a correct word score was calculated for each of these letter-number pairs: 1.0 if both the letter and number had been correctly entered; 0.5 if either the letter or the number had been correctly entered; 0.0 for an incorrectly-entered letter-number pair.

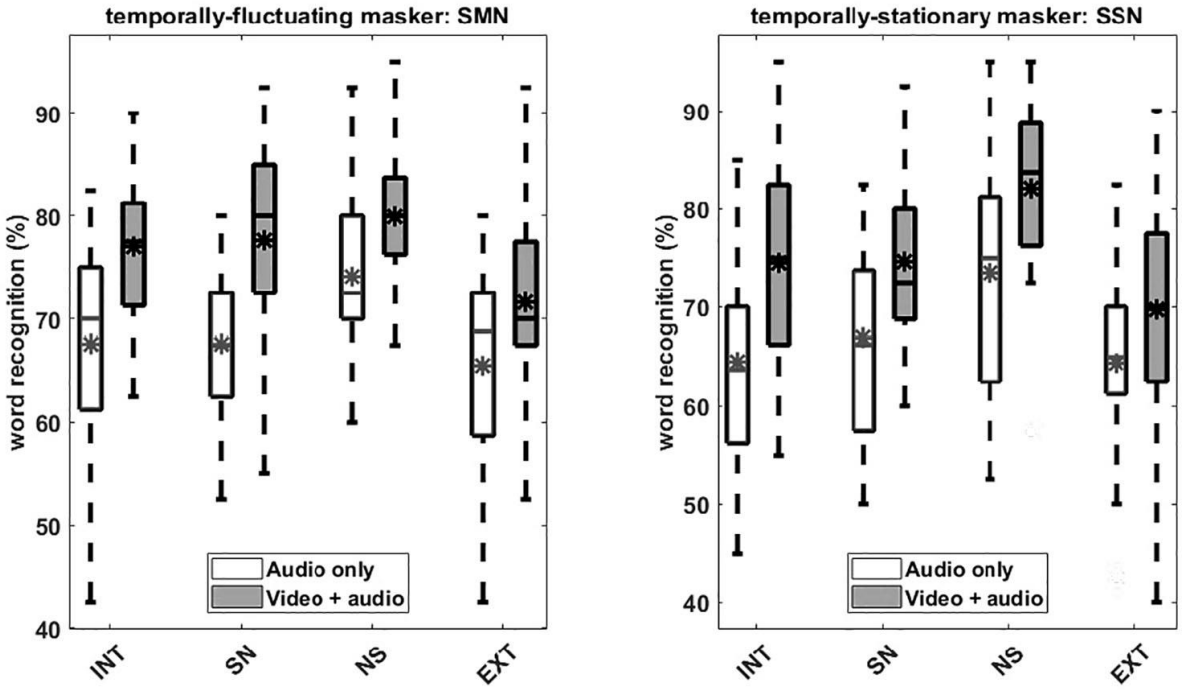


Figure 2. Comparison of overall word recognition percentages across all participants and all 16 combinations of variables in the 'Speech-To-Screen' listening experiment. The left figure refers to target speech in Speech-Modulated Noise (SMN); the right figure refers to target speech in Speech-Shaped Noise (SSN). Binaural auralization positions of the target speech and masking noise: 'INT', 'SN', 'NS', 'EXT'. Mean represented by '*'.

Scores were summed for each participant for each of the 16 combinations of variables and then divided by 20 to calculate the average performance. These results are presented as boxplots in Figure 1, where the word recognition percentages are proxies for speech intelligibility.

The box plots and frequency analysis confirmed that the data for each of the 16 combinations of the 3 independent variables fulfil the criteria for normal distribution. Therefore, since all participants were tested against all the conditions in this listening experiment, a 3-way Repeated Measures ANOVA was used to compare between the independent variables. Mauchly's test indicated that the assumption of sphericity had not been violated.

Two strong main effects were observed: the binaural auralization ($F(3,57) = 22.179$, $p < .0001$, $\eta_p^2 = .805$), and the audio-visual playback condition (video on/off) ($F(1,19) = 25.228$, $p < .0001$, $\eta_p^2 = .570$), suggesting that both the auralization method and the presence of visual cues independently significantly affected the participants' performance in this task. There were found to be no significant two- or three-way interaction effects.

Post-hoc pairwise comparisons were conducted with the Bonferroni correction applied to reduce the probability of a cumulative Type I error to $< 0.05$. Several significant results were determined.

Within the main effect of the binaural auralization, there were significant differences between the means of the word recognition scores for position 'NS' – masking noise internalized at the

headphones and target speech externalized at the screen – versus the three other binaural auralization positions ($p < 0.001$).

| | INT | SN | EXT |
|---|---|---|---|
| **NS** | +9.17% | +7.98% | +14.15% |

Table 2. Ratio gain improvements in speech intelligibility when results for binaural auralization position 'NS' are compared with the 3 other spatialization positions.

Additionally there was a significant difference between the means of the word recognition scores for binaural auralization 'SN' – speech internalized at the headphones and masking noise externalized at the screen – compared to 'EXT' – both target speech and masking noise externalized at the screen ($p < 0.005$). The ratio gain improvement in speech intelligibility for 'SN' compared to 'EXT' was +5.71%.

Within the main effect of the audio-visual playback condition, the difference between the means of the word recognition scores for the video-on condition versus the audio-only condition was significant ($p < 0.001$). There was a ratio gain improvement in speech intelligibility of +11.57% when participants were able to see the videos of the speakers.

## 4 Discussion

The results from the 3-way Repeated Measures ANOVA and post-hoc pairwise comparisons indicate that both the binaural auralization and video (on/off) variables independently had significant effects on the results of this listening experiment.

### 4.1 Binaural auralization

As per our hypothesis, 'NS' – internalizing the masking noise in the headphones whilst externalizing the speech at the screen – results in the most significant improvement in speech intelligibility relative to all three other binaural auralization positions. The improvement found over the other cases where speech and noise were co-located – 'EXT' and 'INT' – was as expected. This is consistent with the hypothesis that spatial separation creates a release of speech from masking and so improves intelligibility.

The improvement of speech intelligibility under condition 'NS' compared to 'SN' – target speech internalized at the headphones and masking noise externalized at the screen – implies that the release of speech from masking is not sufficient to explain all results. There are several possible explanations.

For example, subjects may have subconsciously perceived the 'SN' binaural auralization as a less plausible reproduction, because the speech is not localized at the screen. Furthermore, since in this experiment the externalized speech was reproduced as a point source (for 'NS') versus the externalized masking noise as a diffuse source (for 'SN') it is possible that there were different degrees of externalization for the different types of sounds, even though head-tracking for dynamic binaural synthesis was used in order to aid externalization.

### 4.2 Theoretical spectral masking analysis

We hypothesize that there may be differences in the frequency response depending on the binaural auralization condition.

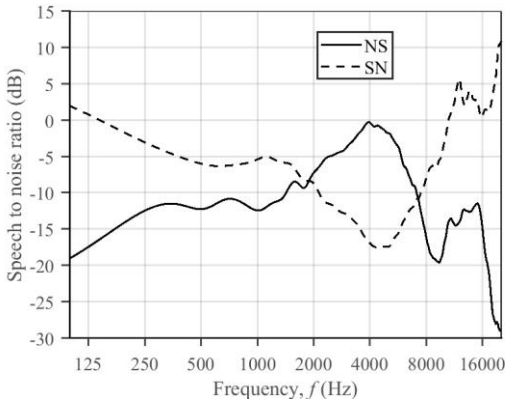


Figure 3. The predicted frequency response of target speech relative to masking noise for the -9 dB SNR temporally-stationary (SSN) masker condition for both 'NS' and 'SN' binaural auralization conditions. Externalized signals have been obtained for the direct signals only (BRIRs truncated before first significant room reflection) using the front-facing BRIRs (i.e. no head rotation).

Analysis with the BRIRs used indicates that for the case of 'NS' – external speech and internal masker – there is more energy in the speech relative to the masking noise approximately between 2.0-7.0 kHz. This is the frequency range within which the most consonant content of speech is contained [22][23]. Conversely, for 'SN' – internal speech and external masker – analysis of the BRIRs suggest that there is more energy in the speech relative to the masking noise below 2.0 kHz and above 7.0 kHz. Energy below 2.0 kHz corresponds to important speech information such as F0, harmonics, and some formants (F1 and F2), which are related to the intelligibility of vowels [24]. Energy above 7.0 kHz, however, contains almost no speech information [25] other than for some voiceless fricatives [26][27][28], and so is mostly redundant to improving intelligibility. This would seem to support the results from the experiment, that is to say, binaural auralization condition 'NS' producing a greater improvement to speech intelligibility than 'SN'.

As the playback signals were recorded during the listening experiment, one further objective analysis which could be conducted as future work would be to pass the combined target speech and masking noise signals from our experiment through an intelligibility model. This would allow the modelled speech intelligibility for each condition to be compared against the experimental results.

### 4.3 Video on/off

The improvement to speech intelligibility with video cues is as expected. Despite the small-screen-sized formatting of the GRID video footage for this experiment, several participants anecdotally mentioned that they had either actively or passively used lip reading during the video-on sessions. A study conducted by Lan et al. (2009) [29], which also used the GRID speech corpus, has implied that it is not solely the shape of the mouth that conveys information for lip reading, but also the visual of the inner part of the mouth.

## 5 Conclusions

This study investigated a new approach to improving the speech intelligibility of audio reproduced over headphones. The application is to future technological developments for small-screen devices such as mobile phones and tablets. The most significant improvement was gained by spatially separating the target speech from the masking noise by rendering one externally and one internally.

A psychoacoustics listening experiment was conducted in which participants listened via headphones to speech sentences in energetic masking noise, and were tasked with correctly identifying target words within each sentence. The experiment examined: 4 binaural auralization positions of all combinations of internalized / externalized speech and masker, 2 video conditions (on / off), and 2 types of energetic masking noise (speech-shaped noise / speech-modulated noise). A 3-way Repeated Measures ANOVA and post-hoc pairwise comparisons revealed that the 'NS' binaural auralization – masking noise internalized at the headphones and target speech externalized at the screen – and the video-on condition were the most significant in improving speech intelligibility. The improvement in speech intelligibility from the binaural processing (+9.2%) is similar to that achieved when lip reading is possible (+11.6%). Plausible explanations for the results of this experiment include the effect of spatial release from masking, differing degrees of externalization of different sounds (point source versus diffuse source), and the differences in frequency response depending on the binaural auralization condition.

## Acknowledgements

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

## References


[1] J. McNally and B. Harrington, “How millennials and teens consume mobile video,” in *TVX*, 2017, pp. 31–39.

[2] J. M. Rigby, D. P. Brumby, S. J. J. Gould, and A. L. Cox, “Media multitasking at home: a video observation study of concurrent TV and mobile device usage,” in *TVX*, 2017, pp. 1–8.

[3] T. Walton, M. Evans, D. Kirk, and F. Melchior, “Does environmental noise influence preference of background-foreground audio balance?,” in *Audio Engineering Society 141st Convention* , 2016, pp. 1–10.

[4] British Broadcasting Corporation, “Compelling TV with good audio,” *BBC Academy*, 2014. [Online]. Available: http://www.bbc.co.uk/academy/en/articles/art20140303161136514. [Accessed: 23-Nov-2017].

[5] W. Brimijoin, A. Boyd, and M. Akeroyd, “The contribution of head movement to the externalization and internalization of sounds,” *PLoS One*, vol. 8, no. 12, p. e83068, 2013.

[6] J. Swaminathan, C. R. Mason, T. M. Streeter, V. Best, E. Roverud, and G. Kidd, “Role of binaural temporal fine structure and envelope cues in cocktail-party listening,” *J. Neurosci.*, vol. 36, no. 31, pp. 8250–8257, 2016.

[7] A. Bronkhorst, “The Cocktail Party Phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” *Acust. - acta Acust.*, vol. 86, no. 1, p. 1170128, 2000.

[8] A. Westermann and J. M. Buchholz, “Release from masking through spatial separation in distance in hearing impaired listeners,” in *Proceedings of Meetings on Acoustics*, 2013, vol. 19, no. 1.

[9] A. Plail and B. M. Fazenda, “On the subjective nature of binaural externalisation,” *Proc. Inst. Acoust.*, vol. 35, no. 2, p. 12, 2013.

[10] M. Armstrong, “From Clean Audio to Object Based Broadcasting,” *BBC Res. Dev. White Pap.*, vol. WHP 324, pp. 1–23, 2016.

[11] H. Fuchs, S. Tuff, and C. Bustad, “Dialogue Enhancement - technology and experiment (EBU Technical Review),” Geneva, 2012.

[12] C. Bustad, L. Mossberg, and H. Wessman, “Sveriges Radios horbarhetsprojekt.” Sveriges Radio, Stockholm, pp. 1–5, 2012.

[13] R. Wallvide, “Slutrapport projektgenomforande - Sveriges Radio,” Stockholm, 2013.

[14] J. Barker and M. Cooke, “Modelling speaker intelligibility in noise,” *Speech Commun.*, vol. 49, pp. 402–417, 2007.

[15] 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,” *J. Acoust. Soc. Am.*, vol. 61, pp. 1337–1351, 1977.

[16] A. R. Bradlow, G. M. Torretta, and D. B. Pisoni, “Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics,” *Speech Commun.*, vol. 20, pp. 255–272, 1996.

[17] International telecommunications Union, *Recommendation ITU-R BS.1116-1 Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. The ITU Radiocommunication Assembly, 1994,

pp. 1–26.

[18] D. Satongar, C. Pike, and Y. Lam, "SBSBRIR - Salford-BBC Spatially-sampled Binaural Room Impulse Responses," 2014. [Online]. Available: http://www.bbc.co.uk/rd/publications/sbsbrir. [Accessed: 07-Nov-2017].

[19] R. Hughes and C. Pike, "(poster) Headphone simulation of 3D spatial audio systems in different listening environments," in *BBC Sound Now and Next*, 2015.

[20] J. Barker, M. Cooke, S. Cunningham, and X. Shao, "The GRID audio visual sentence corpus," *Sheffield University*, 2013. [Online]. Available: http://spandh.dcs.shef.ac.uk/gridcorpus/. [Accessed: 10-Feb-2017].

[21] C. Pike, F. Melchior, and A. Tew, "Descriptive analysis of binaural rendering with virtual loudspeakers using a rate-all-that-apply approach," in *Audio Engineering Society Conference on Headphone Technology*, 2016, pp. 1–8.

[22] T. J. Edwards, "Multiple features analysis of intervocalic English plosives," *J. Acoust. Soc. Am.*, vol. 69, pp. 535–547, 1981.

[23] J. Coleman, "Acoustic structure of consonants," 2017. [Online]. Available: http://www.phon.ox.ac.uk/jcoleman/consonant_acoustics.htm. [Accessed: 13-Mar-2018].

[24] J. C. Catford, "The Cardinal Vowels," in *A Practical Introduction To Phonetics* , 2nd ed., Oxford: Oxfod University Press, 2001, pp. 113–162.

[25] K. Sharma, A. Krishna, and N. U. Cholayya, "Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology," *EURASIP J. Adv. Signal Process.*, vol. 2007, pp. 1–9, 2006.

[26] B. B. Monson, A. J. Lotto, and B. H. Story, "Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives," *J. Acoust. Soc. Am.*, vol. 132, p. 1754, 2012.

[27] A. Jongman, "Acoustic characteristics of English fricatives," *J. Acoust. Soc. Am.*, vol. 108, p. 1252, 2000.

[28] K. Maniwa, A. Jongman, and T. Wade, "Acoustic characteristics of clearly spoken English fricatives," *J. Acoust. Soc. Am.*, vol. 125, p. 3692, 2009.

[29] Y. Lan, R. Harvey, B. Theobald, E. J. Ong, and R. Bowden, "Comparing visual features for lip reading," in *AVSP*, 2009.