Speech-in-noise performances in virtual cocktail party using different non-individual Head Related Transfer Functions

Lorenzo Picinali(1), Maria Cuevas Rodriguez(2), Daniel Gonzalez Toledo(2), Arcadio Reyes Lecuona(2)

(1)Imperial College London, UK, l.picinali@imperial.ac.uk
(2)University of Malaga, Spain, mariacuevas@uma.es
(3)University of Malaga, Spain, dgonzalezt@uma.es
(4)University of Malaga, Spain, areyes@uma.es

Abstract

It is widely accepted that, within the binaural spatialisation domain, the choice of the Head Related Transfer Functions (HRTFs) can have an impact on localisation accuracy and, more in general, realism and sound sources externalisation. The impact of the HRTF choice on speech-in-noise performances in cocktail party scenarios has though not yet been investigated in depth. Within a binaurally-rendered virtual environment, Speech Reception Thresholds (SRTs) with frontal target speaker and lateral noise maskers were measured for 22 subjects several times across different sessions, and using different HRTFs. Results show that for the majority of the tested subjects, significant differences could be found between the SRTs measured using different HRTFs. Furthermore, the HRTFs leading to better or worse SRT performances were not the same across the subjects, indicating that the choice of the HRTF can indeed have an impact on speech-in-noise performances within the tested conditions. These results suggest that when testing speech-in-noise performances within binaurally-rendered virtual environments, the choice of the HRTF should be carefully considered. Furthermore a recommendation should be made for future modelling of the speech-in-noise perception mechanisms to include monoaural spectral cues in addition to interaural differences.

Keywords: Speech, Binaural, HRTF

1 INTRODUCTION

Binaural spatialisation is a technique which allows the creation of three-dimensional soundscapes through a simple pair of headphones, and is widely used in interactive audio and Virtual Reality (VR) applications. The most traditional implementation of this technique is based on the convolution between a mono signal and a Head Related Transfer Function (HRTF), generally measured from an individual's ears or from a dummy head microphone. Previous studies showed that spatialising sounds using an HRTF different from someone's own can have an impact on localisation accuracy, for example in tasks such as front/back and up/down discrimination [1]. Using a non-individual HRTF could also have an impact on realism, immersiveness and other qualitative judgements of the simulated sound fields, even though high variability has been observed between expert and naive listeners, resulting in low repeatability of the evaluation across different sessions [2].

It is known that a link exists between sound localisation and speech-in-noise perception, more specifically related with the advantages gained by being able to spatially separate the speech and noise sources [3]. This phenomenon is known as Spatial Release from Masking (SRM), and is strongly associated with the so-called Cocktail Party effect [4], i.e. our ability to focus on a single talker or conversation in a noisy room.

The impact of using different non-individual HRTFs when measuring speech-in-noise performances in virtual (binaurally spatialised) cocktail party scenarios has not yet been explored extensively. This article briefly outlines a study focussing on the measurement Speech Reception Threshold (SRT) with frontal speech target and lateral maskers, all rendered through a pair of headphones using the binaural spatialisation technique. Different non-individual HRTFs have been used for each subject, and a comparison between SRTs obtained with the different filters, across different sessions, has been carried out.

2 METHOD

Using the 3D Tune-In Toolkit binaural spatialiser [5], a virtual scenario with a frontal target speech source and two lateral (±90°) masking sources was created. Bisyllabic words were used for the frontal speech, and uncorrelated babble noise for the maskers. SRT in noise corresponding to the 50% correct repetition rate for the words was recorded through a simple up-down 2 dB step adaptive procedure [6]. Seven different measured HRTFs [7], plus a synthetic one created through a simple one-pole one-zero model (simulating Interaural Level Differences - ILDs) and a delay (simulating Interaural Time Differences - ITDs), were compared.

22 participants performed the test. Every individual carried out 20 sessions at different times across several days, each one consisting in eight SRT measurements (7 measured HRTFs + 1 synthetic) in random order.

3 RESULTS

A preliminary analysis has been carried out looking at each subject individually, and averaging the SRT values obtained in the various sessions separately for each HRTF. The first significant result is that for almost all the individuals who took part to the test the HRTF that resulted in the worst SRT was the synthetic one. This difference was statistically significant in over 80% of the cases. Looking at the comparison between the 7 measured HRTFs, for 9 subjects statistically significant differences could be observed between the measured SRTs, indicating that in those cases the choice of the HRTF had a significant impact on the outcome measure. For the remaining 13 subjects differences in SRTs between the HRTFs were observed, but these were not statistically significant. Finally, it is important to note that the HRTFs leading to better or worse SRT performances were not the same across the subjects, and that almost every HRTF lead at least once to the best recorded SRT.

Considering the comparisons between the synthetic and the measured HRTFs, the result seem to indicate that the spectral monoaural features of the measured HRTFs, which were not modelled in the synthetic one, are indeed relevant when trying to understand speech in noise in a cocktail party scenario. The comparison between the different measured HRTFs then outlines that for certain subjects there is a significant effect of the HRTF choice (not always the same one across subjects) on the measured SRT.

These results suggest that when testing speech-in-noise performances within binaurally-rendered virtual cocktail party scenarios, the choice of the HRTF could have an impact on the outcome SRT measure, and should therefore be carefully considered. Furthermore a recommendation should be made for future modelling of the speech-in-noise perception mechanisms to include monoaural spectral cues in addition to interaural differences.

REFERENCES

[1] Begault, D.R.; Wenzel, E.M.; Anderson, M.R. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. Journal of the Audio Engineering Society. 2001 Oct 1;49(10). pp 904-16.

[2] Andreopoulou, A.; Katz, B. Investigation on subjective HRTF rating repeatability. InAudio Engineering Society Convention 140 2016 May 26. Audio Engineering Society.

[3] Dirks, D.D.; Wilson, R.H. The effect of spatially separated sound sources on speech intelligibility. Journal of Speech and Hearing Research. 1969 Mar;12(1). pp 5-38.

[4] Cherry EC. Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America. 1953 Sep;25(5). pp 975-9.

[5] Cuevas-Rodriguez, M.; Picinali, L.; Gonzalez-Toledo, D.; Garre, C.; de la Rubia-Cuestas, E.; Molina-Tanco, L.; Reyes-Lecuona, A. 3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation. PloS one, 2019, 14(3), e0211899.

[6] Levitt, H. Adaptive testing in audiology. Scandinavian audiology. Supplementum. 1978(6), pp 241-91.

[7] Katz BF, Parseihian G. Perceptually based head-related transfer function database optimization. The Journal of the Acoustical Society of America. 2012 Feb 13;131(2), pp 99-105.