TECNIACUSTICA MURCIA · 2014

45º CONGRESO ESPAÑOL DE ACÚSTICA
8º CONGRESO IBÉRICO DE ACÚSTICA
EUROPEAN SYMPOSIUM ON SMART CITIES AND ENVIRONMENTAL ACOUSTICS

ESTUDIO DE LA INFLUENCIA DE LA AURALIZACIÓN EN LA INTELIGIBILIDAD Y LA INMERSIÓN ACÚSTICA EN SISTEMAS DE MULTICONFERENCIA UTILIZANDO SONIDO BINAURAL

REFERENCIA PACS: 43.60.Dh

Aguilera, Emanuel; Lopez, Jose Javier; Gutierrez-Parera, Pablo

Universidad Politécnica de Valencia
ITEAM, Instituto de Telecomunicaciones y Aplicaciones Multimedia
Camino de Vera s/n, Edificio 8G, acceso D
46022 Valencia (España)
Tel: +34 963 877 007 (Ext. 73008) Fax: +34 963 879 583
emagmar@iteam.upv.es, jjlopez@dcom.upv.es, pabgupa@iteam.upv.es

ABSTRACT

Spatial audio improves intelligibility and immersion of conventional teleconferencing systems leveraging the ability of the human auditory system to locate, separate and understand multiple speakers when they talk simultaneously. Moreover, it is possible to increase this immersion if auralization is introduced to simulate the virtual room, although too much auralization can affect speech intelligibility. In this paper we study the influence of auralization on speech intelligibility and immersion in a multi-party teleconference system developed for mobile devices using binaural audio. Different subjective experiments evaluate the influence of the early echoes, the late reverberation, the room size and other details related to binaural processing.

RESUMEN

El audio espacial mejora la inteligibilidad e inmersión de los sistemas tradicionales de teleconferencia aprovechando las capacidades del sistema auditivo humano para localizar, separar y entender a varias personas hablando simultáneamente. Además, es posible aumentar esta inmersión añadiendo una auralización que simule la sala virtual, aunque si se introduce en exceso puede afectar a la inteligibilidad. En este trabajo se estudia la influencia de la auralización en la inteligibilidad e inmersión en un sistema de multiconferencia con audio binaural desarrollado para dispositivos móviles. Varios experimentos subjetivos evalúan la influencia de las primeras reflexiones, la reverberación tardía, el tamaño de la sala y otros detalles del procesado binaural.

1. INTRODUCCIÓN

Los servicios de multiconferencia tienen un importante papel en la interacción social, proporcionando ventajas incuestionables hoy en día. Son muy útiles, especialmente en el ámbito de los negocios, ya que ahorran mucho tiempo y gastos en viajes [1,2]. Por esta razón desde hace mucho tiempo las empresas han estado usando estos servicios a través de la red de telefonía fija.

Un terminal móvil también puede usar servicios de multiconferencia como los ofrecidos por las operadoras a través de la telefonía fija [3]. La primera y más obvia ventaja es que los usuarios empresariales ganan movilidad y no están obligados a permanecer en sus oficinas. Del mismo modo, con un sistema de acceso y tarifas adecuado, este servicio puede ser atractivo para usuarios particulares (no empresariales) que podrían realizar reuniones de audio con amigos y familiares desde cualquier lugar. Además, hoy en día los terminales móviles tienen muchas más prestaciones que los terminales de telefonía fija, que han sido relegados prácticamente solo a comunicaciones de voz.

Otra ventaja que adquiere actualmente un fuerte interés comercial es la idea de aprovechar las capacidades de los terminales móviles para que las multiconferencias sean más realistas. Una de estas mejoras está relacionada con la incorporación de audio espacial a la conferencia [3,4].

En anteriores trabajos se ha demostrado que el audio espacial (estéreo o binaural) es preferido por los usuarios más que el audio monoaural no espacial [4,5], mejorando la discriminación del hablante [5,6] y la inteligibilidad de las conversaciones ya que se aprovecha la habilidad del sistema auditivo humano de prestar una atención selectiva a los sonidos de una determinada dirección (“*cocktail party effect”*).

Por otra parte, la incorporación del efecto de sala (auralización) en la audioconferencia produce resultados controvertidos según estudios anteriores [7], ya que, aunque incrementa la percepción de distancia, degrada la dirección de la fuente de audio y su inteligibilidad.

En este trabajo se propone un sistema nuevo de multiconferencia inmersiva para dispositivos móviles (teléfonos inteligentes y tabletas) que mejora sustancialmente la identificación e inteligibilidad de los participantes. Esto se logra por medio de la reproducción de sonido binaural con auriculares para localizar a los diferentes hablantes en una sala de reuniones virtual y usando una gran pantalla táctil para que el usuario mueva y coloque interactivamente a los participantes en cualquier posición espacial. Este sistema desarrollado por nosotros incluye la aplicación del terminal móvil y el servidor que gestiona las comunicaciones.

Usando este sistema se van a analizar los efectos de añadir diferentes niveles de auralización a la audioconferencia con pruebas subjetivas a un grupo de participantes. La idea es investigar individualmente la influencia de las primeras reflexiones, la reverberación tardía, el tamaño de la sala y otros detalles que se utilizan en los métodos clásicos de auralización.

El resto del artículo está organizado de la siguiente manera: el apartado 2 describe el desarrollo del software utilizado en los experimentos y la tecnología relacionada, incluyendo los algoritmos de auralización utilizados. La sección 3 describe el experimento subjetivo y sus objetivos. El apartado 4 expone los resultados así como su análisis y discusión. Por último la sección 5 presenta las conclusiones de este trabajo.

2. IMPLEMENTACIÓN

Para la realización del test se ha utilizado una aplicación de multiconferencia desarrollada en el grupo de investigación y que funciona sobre teléfonos inteligentes (iPhone y Android) y tabletas (iPad). El software permite a los participantes controlar fácilmente su uso por medio de una

interfaz gráfica y táctil. Esta aplicación utiliza audio binaural usando un modelo simple de HRTF (*Head-Related Transfer Function*). Cada usuario puede mover y colocar virtualmente a los otros participantes alrededor de su propia posición en la mitad del borde inferior (ver figura 8). Se puede encontrar una descripción completa de la aplicación en [8].

2.1. Sonido binaural

En la práctica, las funciones HRIR (*Head-Related Impulse Response*) tienen una longitud entre 128 y 512 muestras. Convolucionar cada señal mono con las dos HRIR supone un importante coste computacional. A pesar de que estas convoluciones se puedan realizar en el dominio frecuencial con multiplicaciones FFT (usando algoritmos *overlap-add* u *overlap-save*) de un modo eficiente, este proceso sigue teniendo un gran coste. Y aunque hoy en día los teléfonos inteligentes tienen suficiente potencia para realizar en tiempo real estos algoritmos de filtrado, se tendría que utilizar un gran parte de los recursos computacionales en esta etapa de procesado, reduciéndose también la duración de la batería.

Por estas razones se ha pensado en una implementación eficiente de la HRTF para esta aplicación. Para simplificar mucho el coste computacional se ha implementado una aproximación de la HRTF que proporciona una síntesis suficientemente buena.

La implementación de la HRTF se ha dividido en dos partes. Primero se ha implementado la ILD (*Interaural Level Difference*) con dos filtros IIR (*Infinite Impulse Response*). Segundo, la ITD (*Interaural Time Difference*) se ha conseguido añadiendo un retardo temporal entre las señales del oído izquierdo y derecho. El diseño de los filtros ILD IIR se ha realizado siguiendo un procedimiento similar a [9], donde los autores obtuvieron un modelo HRTF estándar promediando las respuestas de una base de datos de respuestas HRTF reales. Usando la respuesta promediada, se ha ajustado un filtro IIR de 6º orden paramétrico para cada dirección azimut. Se pueden encontrar más detalles de esta implementación en [8].

2.2. Auralización

La aplicación simula que los participantes están en una habitación virtual cuyo tamaño se extiende por la pantalla. El algoritmo de auralización tiene que simular las reflexiones y otros efectos acústicos que suceden en una sala teniendo en cuenta la posición de los participantes y del oyente. Existen muchos métodos precisos para modelar la respuesta al impulso de una sala, RIR (*Room Impulse Response*), entre dos puntos de la sala, muchos de ellos basados en modelos físicos de la propagación del sonido. Para esta aplicación no es necesario un método demasiado preciso, basta con uno que produzca una sensación realista.

La RIR se describe generalmente como la composición de una señal directa, primeras reflexiones y una cola de reverberación. Nuestra aplicación calcula las primeras reflexiones usando el método de la imagen. Se tienen en cuenta los cuatro ecos que proceden de las reflexiones de primer orden de las paredes laterales. Adicionalmente la aplicación cuenta con la posibilidad de procesar las reflexiones binauralmente (considerando la dirección de llegada) o monoauralmente. Para la cola de reverberación se utiliza un reverberador clásico de Schroeder [10].

3. DESCRIPCIÓN DEL EXPERIMENTO

El objetivo del experimento es investigar individualmente la influencia de las partes que conforman la RIR (primeras reflexiones y reverberación tardía), el tamaño de la sala virtual y también el efecto de usar primeras reflexiones filtradas binauralmente o simplemente monoaurales.

El test se ha realizado con nuestra aplicación de multiconferencia sobre tabletas iPad usando auriculares Sennheiser HD 439. Estos auriculares son de tipo cerrado para tener un buen aislamiento del ruido externo y circumaurales porque producen menos fatiga y son más cómodos de llevar puestos. En el experimento han participado un total de 10 personas (5 hombres y 5 mujeres), todos sin problemas de audición con edades comprendidas entre los 22 y 40 años, siendo estudiantes y personal de diferentes grupos de investigación de la universidad. Los participantes tenían que escuchar con este sistema una conversación entre tres personas grabada previamente pudiendo elegir libremente la posición de los hablantes. Por ello se les invito a mover interactivamente los avatares con la pantalla táctil a diferentes posiciones de la sala virtual todo lo que quisieran durante el test para que notaran como la percepción espacial cambia con la posición y con las diferentes opciones de auralización ofrecidas.

Dado el gran número de parámetros a probar se decidió dividir el test en dos fases. En la primera se analizaron dos tipos de primeras reflexiones: filtradas binauralmente y monoauralmente. El primer tipo utiliza ligeramente más potencia de cálculo que el segundo y es teóricamente más realista, pero la diferencia es tan sutil que era interesante probar si los oyentes podían detectarlo, y al mismo tiempo saber sus preferencias. Además, este análisis previo permite reducir la cantidad de combinaciones a evaluar en la segunda fase del test.

En esta primera fase la aplicación muestra un botón en la pantalla para cambiar de forma ciega entre los dos diferentes procesados de las primeras reflexiones A/B (test de escala nominal directa [11]). Los participantes tuvieron que rellenar un formulario contestando la pregunta *"¿Cuál de los dos casos produce mayor sensación de realismo y posición?"* y también se les pidió cuantificar la diferencia que percibieron en una escala de 1 (casi iguales) a 5 (totalmente diferentes).

De acuerdo a los resultados de este primer test (analizados en la siguiente sección), se eligió el procesado binaural de las primeras reflexiones para la segunda fase del experimento. En el segundo test los participantes podían cambiar entre seis casos diferentes de auralización que combinan las primeras reflexiones, la reverberación tardía y el tamaño de la sala. Las combinaciones son:

1. Señal directa (sin primeras reflexiones ni reverberación). {SD}
2. Señal directa + primeras reflexiones sala pequeña (5x3.75m). {PR(P)}
3. Señal directa + primeras reflexiones sala grande (10x7.5m). {PR(G)}
4. Señal directa + primeras reflexiones sala pequeña (5x3.75m)+reverberación. {PR(P)+R}
5. Señal directa + primeras reflexiones sala grande (10x7.5m). {PR(G)+R}
6. Señal directa + reverberación. {R}

Usando la misma conversación grabada del primer test, los participantes podían mover libremente los avatares y cambiar en cualquier momento entre los diferentes casos de forma ciega, figura 8. La aplicación cambia los algoritmos de auralización en tiempo real según el botón que vaya seleccionando el usuario. Los botones se asignan aleatoriamente a cada caso de auralización para cada participante para que no influya su orden.

Se les pidió a los participantes evaluar 4 parámetros relacionados con la auralización:

1. Inteligibilidad: *¿Entiendes la conversación fácilmente? ¿Cuánto?*
2. Distancia: *¿Tienes sensación de distancia cuando mueves a los hablantes? ¿Cuánta?*
3. Inmersión: *¿Te sientes inmerso en un entorno o sala real? ¿Cuánto?*
4. Preferencia general: *¿Cuánto te gusta cada opción? Ordénalas según tu preferencia.*

Para cada parámetro los participantes ordenaron los 6 diferentes casos de auralización, de 0 (bajo) a 5 (alto) (test de escala ordinal directa [11]) rellenando una tabla de 4x6 casillas.

4. RESULTADOS Y DISCUSIÓN

4.1. Resultados del primer experimento

Los resultados de la primera parte del test fueron algo sorprendentes. Casi al mismo número de participantes les pareció que las primeras reflexiones monoaurales (40%) producían mayor sensación de realismo y posición que las reflexiones procesadas binauralmente (60%) que en teoría son más realistas. En cualquier caso encontraron muy difícil decidir entre ambos casos porque eran muy parecidos (una media de 2 en una escala de disimilitud de 1 a 5), tabla I.

Primeras reflexiones	Preferencia de los participantes	Diferencia percibida (1-5)
Monaural	40 %	1.75
Binaural	60 %	2

Tabla I. Resultados de la primera parte del test

Teniendo en cuenta estos resultados se ha continuado la siguiente parte del test aplicando las primeras reflexiones procesadas binauralmente de forma que el número de combinaciones de parámetros de auralización se reducen a una pequeña cantidad de casos de estudio evitado así casos demasiado parecidos.

4.2. Resultados del segundo experimento

En la segunda parte del test se han evaluado 4 cualidades (inteligibilidad, percepción de distancia, sensación de inmersión y preferencia general) para seis casos de auralización. Los participantes ordenaron estas cualidades con una puntuación de 0 (el peor caso) a 5 (el mejor caso). Para cada caso y característica se ha calculado la puntuación media y el intervalo de confianza al 95 % (IC).

En cuanto a la inteligibilidad, como muestra la figura 1, el mejor resultado se ha obtenido para el caso de únicamente señal directa (SD) y los participantes encontraron como la más difícil de entender la combinación de primeras reflexiones y reverberación en una sala grande (PR(G)+R).

La evaluación de la percepción de distancia (figura 2) indica que la combinación preferida es la señal directa únicamente con reverberación (R), siendo el caso de solo señal directa el segundo en preferencia. Los casos de auralización que incluyen primeras reflexiones no consiguen gran percepción de distancia y sus intervalos de confianza indican que esta percepción es muy parecida en esos casos.

Cuando se les pregunta por la inmersión (figura 3), los participantes prefieren la señal directa (SD) seguido de los casos con primeras reflexiones. La reverberación no tiene una buena valoración para la sensación de inmersión.

Sorprendentemente, en la preferencia general (figura 4), la señal directa (SD) es claramente el método preferido para un sistema de multiconferencia y el peor valorado es el que incluye primeras reflexiones y reverberación en una sala grande (PR(G)+R).

4.3. Discusión

Cualidades

En las figuras anteriores se puede observar que la inteligibilidad, inmersión y preferencia general tienen una tendencia similar para cada combinación de auralización pero la percepción de la distancia tiene una tendencia opuesta (con la única excepción del caso de señal directa).

La inteligibilidad obtiene los resultados esperados. Una señal limpia es la preferida si se considera que añadir reflexiones tempranas la ensucia y más todavía si se suma reverberación.

La percepción de distancia es mayor si hay reverberación. Según los resultados, las primeras reflexiones no ayudan a mejorar la percepción de distancia, probablemente porque añaden más energía a la señal, de modo que el menor volumen cuando el hablante se aleja no se percibe bien. Cuando se añade reverberación a cualquiera de los 3 primeros casos, la percepción de la distancia aumenta como se ve en la figura 7.

Los resultados de la sensación de inmersión no han sido tan obvios. Cabría esperar que añadir auralización de sala a la señal incrementaría la sensación de realismo pero los participantes dicen que una señal más limpia les hace sentir más inmersos. Muchos de ellos explicaron que esperaban sentirse hablando en una sala bien acondicionada y no reverberante por eso cualquier tipo de eco o reverberación hacía la experiencia menos realista.

En cuanto a la preferencia general, los participantes prefieren la señal limpia, dando más importancia a la inteligibilidad e inmersión y sin preocuparles tanto la percepción de distancia. Es decir, en este tipo de aplicaciones la sensación realista de distancia no es una característica importante.

Casos

- Señal directa (SD): es la mejor opción para una aplicación de multiconferencia en términos de inteligibilidad, inmersión y preferencia de los usuarios. Incluso para la percepción de distancia es una buena opción porque la amplitud decreciente de la señal se aprecia bien cuando se alejan los avatares.

- Primeras reflexiones (PR(P) y PR(G)): como se ha visto en el primer test, las primeras reflexiones binaurales son ligeramente mejores que las monoaurales en cuanto a realismo y posición. Teniendo en cuenta que el uso de CPU es casi el mismo en ambas técnicas, ésta sería la opción adecuada. Con el segundo test se ha encontrado que los primeros ecos son bastante buenos en inteligibilidad e inmersión y una buena opción final, siendo la sala pequeña mejor que la sala grande (figura 5). Si solo consideráramos la percepción de distancia, usar las primeras reflexiones seria la peor opción, siendo la sala pequeña (PR(P)) peor que la sala grande (PR(G)). Cuando se añade reverberación a las primeras reflexiones, la diferencia entre el comportamiento de la sala pequeña (PR(P)+R) y grande (PR(G)+R) aumenta (figura 6).

- Reverberación (R): para este tipo de aplicación, añadir únicamente reverberación a la señal directa tiene resultados de inteligibilidad, inmersión y preferencia parecidos a añadir solamente las primeras reflexiones (PR). Pero la reverberación es la mejor técnica de auralización para aumentar la percepción de la distancia. La figura 7 muestra como al añadir reverberación a una señal (a la directa o a la que tiene primeras reflexiones) la percepción de la distancia es mejor.

- Primeras reflexiones y reverberación (PR(P)+R y PR(G)+R): es el peor caso (sobre todo con la sala grande) en términos de inteligibilidad, inmersión y preferencia general y no es buena para la percepción de distancia, así que no hay razones para usar estas combinaciones en este tipo de aplicación.

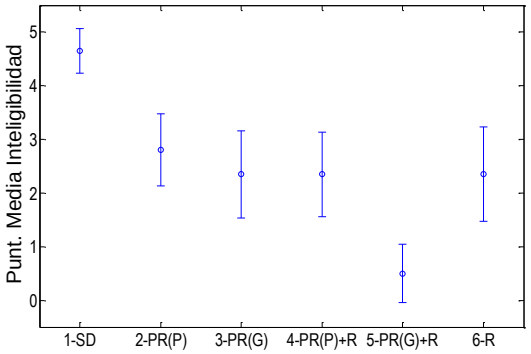

Figura 1. Puntuación inteligibilidad (IC 95%)

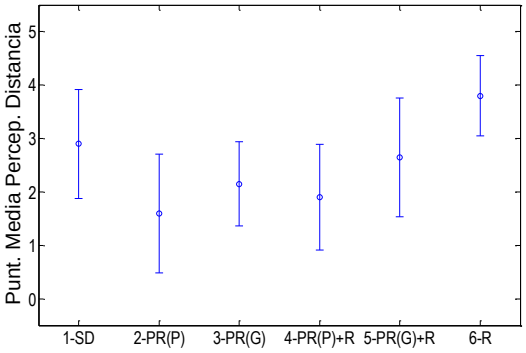

Figura 2. Puntuación percepción distancia (IC 95%)

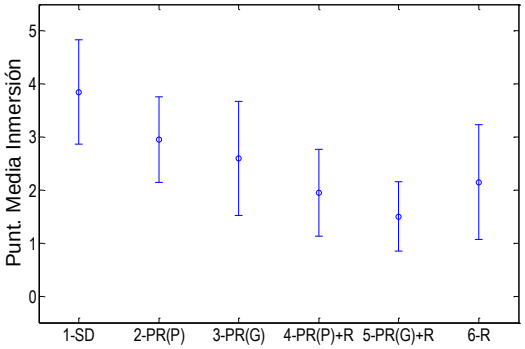

Figura 3. Puntuación inmersión (IC 95%)

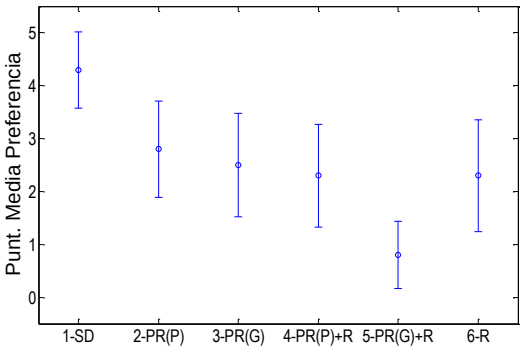

Figura 4. Puntuación preferencia (IC 95%)

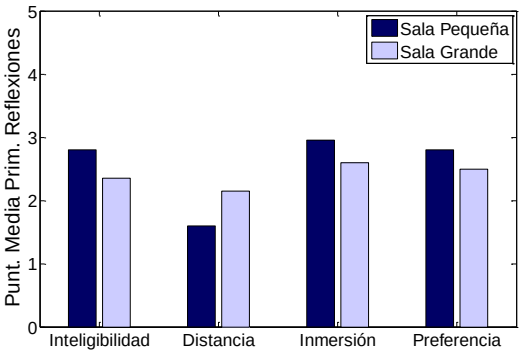

Figura 5. Comparación puntuación según el tamaño de sala al usar primeras reflexiones.

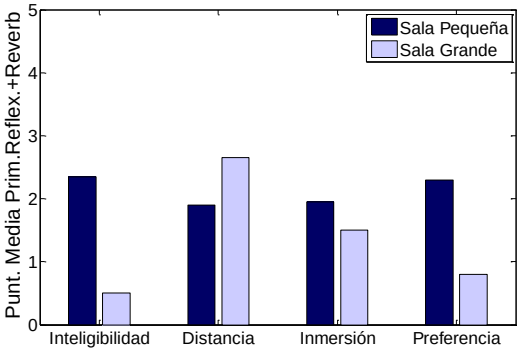

Figura 6. Comparación puntuación según el tamaño de sala al usar primeras reflexiones con reverberación.

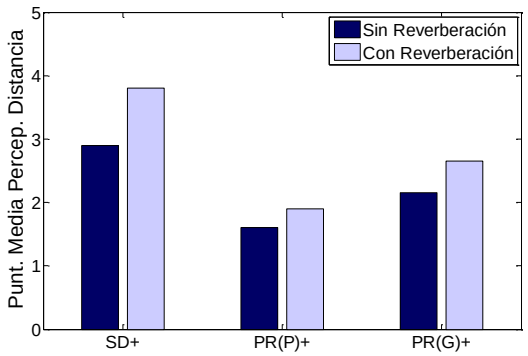

Figura 7. Puntuación distancia con y sin reverberación.

Figura 8. Pantalla principal de la aplicación en tableta.

5. CONCLUSIÓN Y TRABAJO FUTURO

En este trabajo se han realizado varios experimentos subjetivos para evaluar la influencia de las primeras reflexiones, reverberación, el tamaño de la sala y otros detalles cuando se aplica auralización a un sistema de multiconferencia binaural. Aunque se han presentado conclusiones particulares para cada caso en el apartado de discusión, se pueden extraer algunas conclusiones generales. Los participantes prefieren señales más limpias dando mayor importancia a la inteligibilidad e inmersión y sin importarles tanto la percepción de distancia. Utilizar únicamente la señal directa binaural es la mejor opción en cuanto a inteligibilidad e inmersión y es el caso preferido. Incluso para la percepción de la distancia es una buena opción porque se aprecia bien la amplitud decreciente de la señal cuando se alejan los avatares. Por otra parte, un algoritmo de auralización más completo que incluya las primeras reflexiones y reverberación obtiene los peores resultados.

De todos modos, en algunos casos los intervalos de confianza se solapan demasiado para considerar el test completo muy fiable. Como trabajo futuro el experimento se debería realizar con un mayor número de personas para reducir el IC. Además, el nivel de las primeras reflexiones (el coeficiente de reflexión de las paredes de la sala) debería analizarse con un test subjetivo para saber si valores más bajos producen mejores resultados.

AGRADECIMIENTOS

Financiado por el Ministerio Español de Ciencia e Innovación, proyecto TEC2012-37945-C01.

REFERENCIAS

[1] G. M. Olson and J. S. Olson, “Distance matters”, Human-Computer Interaction, Volume 15, pp. 139-178, Lawrence Erlbaym Associates Inc, 2000.

[2] J. Sussman, J. E. Christensen, S. Levy, W. E. Bennett, T. V. Wolf, T. Erickson and W. A. Kellogg, "Rendezvous: Designing a VoIP conference call system", UIST, 2006.

[3] S. Deo, M, Billinghurst, N. Adams and J. Lehikioinen, “Experiments in spatial mobile audio-conferencing”. Proc. of the 4th international conference on mobile technology applications and systems and the 1st international symposium on Computer human interaction in mobile technoloy,vol. 7, ACM Press, pp. 447-451, 2007.

[4] S. Goose, J. Riedlinger and S. Kodlahalli, “3D audio conferencing and archiving services for handheld wireless devices”, Int. Journal of Wireless and Mobile Computing, vol. 1(1), pp. 5-13, 2005

[5] R. Kilgore, M. Chignell and P. Smith, Paul, “Spatialized audioconferencing: what are the benefits?”, Proc. of the 2003 conference of the Centre for Advanced Studies on Collaborative research, pp. 135-144,IBM Press, 2003

[6] J. Baldis, “Effects of spatial audio on memory, comprehension, and preference during desktop conferences”, Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, pp. 166-173, 2001

[7] B. Shinn-Cunningham, “Learning reverberation: Considerations for spatial auditory displays,” in Proceedings of the ICAD, 2000, pp. 126–134.

[8] E. Aguilera, J.J.Lopez, P.Gutierrez, M. Cobos, “An immersive multi-party conferencing system for mobile devices using binaural audio”, Proc. of the AES 55th Conference on Spatial Audio, Helsinki, 2014

[9] J.J. López, M. Cobos, B. Pueo, “Elevation in wave-field Synthesis using HRTF cues”, Acta Acustica, 96 (2), pp. 340-350, 2010.

[10] M. R. Schroeder, ``Natural-sounding artificial reverberation,'' Journal of the Audio Engineering Society, vol. 10, no. 3, pp. 219-223, 1962.

[11] S. Bech and N. Zacharov, “Perceptual Audio Evaluation-Theory, Method and Application”, John Wiley & Sons Ltd., Susex (UK), 2006