SMYTH SVS
HEADPHONE SURROUND MONITORING FOR STUDIOS

Stephen Smyth, Mike Smyth, Steve Cheung

Smyth Research Ltd, Bangor, Northern Ireland, UK

Smyth SVS is an audio virtualisation algorithm for use with standard stereo headphones. Its design goal was simply to reproduce the sound of loudspeakers, in a known environment, with a level of fidelity that would render the headphone presentation indistinguishable from that of the real loudspeakers. To achieve this, room impulse responses from each loudspeaker in a real listening room, captured using miniature microphones placed in the listener's ears, are used to create a personalised virtual listening environment, forming the basis of the SVS algorithm. The algorithm also responds to head-tracking data to maintain the realism and accuracy of the virtual speakers over a limited range of head movements. As implemented in the Smyth Realiser, SVS can be readily used by professionals as a multi-channel surround monitoring tool for recording, post-production and broadcasting. It also has immediate applications in consumer soundtrack playback.

VIRTUALISATION

Normal hearing requires just two audio signals, one from each ear, to sense an entire 3-dimensional external soundscape. Binaural audio reproduction is a technique that attempts to generate these two 'ear' signals in order to re-create a 3-dimensional hearing sensation during playback. Virtualisation, as defined herein, is a binaural reproduction technique that aims to re-create virtual loudspeakers that are indistinguishable from reality. This process requires information about how the loudspeakers, the environment and the listener, all affect the audio signals before they are finally heard - information that can be determined from the impulse responses of the three individual parts.

Binaural Room Impulse Response

When an impulse-type excitation signal is generated by a loudspeaker, the sound waves move out from the speaker and are reflected, diffracted and absorbed by all the surfaces in the room to create a decaying, reverberant sound field. This sound field defines the combined impulse responses of the amplifier, speaker and the room, hereinafter termed the Room Impulse Response (RIR). Similarly, excitation sounds incident on a listener's ear are modified by the filtering effects of the head and outer ears, commonly termed the Head Related Impulse Response (HRIR) or Head Related Transfer Function (HRTF) in the frequency domain. A loudspeaker excitation signal that is both modified by the room and by a listener's ear, as measured at the ear canals, becomes the Binaural Room Impulse Response (BRIR), and contains all the information needed to exactly replicate the sound of the original loudspeaker in the room for the subject listener. During binaural reproduction, the BRIR for each source speaker is convolved with the audio signal for that speaker, and the convolved signals for all the virtual speakers are summed together and finally output as a 2-channel signal for playback over headphones.

BINAURAL REPRODUCTION PROBLEMS

Individualised HRTFs

However a major issue for any accurate binaural reproduction system is the requirement for individualised HRTFs. [1][4][5] It has been commonly accepted that individualisation, whilst necessary for authenticity, is not practical, and that generalised HRTFs are more appropriate for commercial systems. This paper forwards an opposing view, based on the development of our SVS system, that realistic virtualisation requires personalised BRIRs (i.e. individualised HRTFs) if headphone surround monitoring is to become widely accepted.

Externalisation

Another debilitating issue for binaural audio, that is particular to its reproduction over headphones, is the movement of virtual speakers caused by head rotations. This causes auditory confusion for the listener, and quickly ruins the illusion of externalized virtual speakers.

PRIR: personalised BRIR

Smyth SVS attempts to address both of these issues operating within the fairly rigid requirements of sound monitoring in a studio environment. A personal BRIR measurement system and methodology has been designed and fully integrated within the core SVS algorithm. This allows users to readily acquire personalised BRIR measurements (termed PRIR measurements or data) for up to eight active loudspeaker sources within a target monitoring environment. The accuracy with which the PRIR data can be measured directly impacts on the overall accuracy of the SVS virtualisation system.

Dynamic de-rotation of the measured PRIR data is also incorporated in the core algorithm, and a low-cost optical tracking system designed, for use with stereo headphones within a project studio environment.

The current SVS convolution engine, running on a single floating point DSP, is capable of virtualising, and tracking, eight full-bandwidth speakers, in any position, with a maximum convolved reverberation time of 800ms.

HEADPHONE SURROUND MONITORING

Advantages

If virtualisation could be made accurate enough, then headphone monitoring would enjoy some definite advantages over loudspeaker based monitoring.

- lower cost - should be cheaper than loudspeakers + room + room treatment for an equivalent fidelity
- portable - a reference sound room could be brought to the source
- more versatile - different formats, rooms and speakers would all be instantly available
- acoustic isolation - headphones make quiet neighbours compared to speakers

Other, rather less obvious advantages, include:

- centred centre channel - virtual speakers need never be obstructed by a video screen (no need for perforated screens)
- dynamic sweet spot - the sweet spot moves with the user and multiple users can all have identical sweet spots irrespective of seating position (everyone gets to hear the same thing)

However, all these advantages presuppose an adequate level of absolute acoustic accuracy. Therefore the development of SVS was based on the concept of transparent rendition of real loudspeakers in a real listening room.

Other systems have attempted this commercially, notably Lake Technology's TheaterPhone system, Studer's Binaural Room Scanning system [3] developed in collaboration with IRT, and more recently Beyerdynamic's HeadZone system developed with Sonic Emotion. All of these systems use non-individualised binaural data for loudspeaker rendering.

Smyth SVS on the other hand was conceived primarily to facilitate the use of personalized measurements. To achieve this, it was necessary to develop a method of acquiring such data that was both practical and sufficient for transparent rendition. Although this approach would require a level of 'user-customisation' not normally associated with audio playback systems, it was assumed that professional users would willingly undergo the inconvenience of having to take binaural measurements if the resulting accuracy was sufficiently good. Once convinced of its basic accuracy, many of the advantages listed above might then become relevant.

SMYTH SVS

The application of Smyth SVS to headphone surround monitoring is straight forward. The user first makes a set of personalized measurements in the acoustical environment (speakers + room) that they wish to render over headphones. All the routines necessary for this process are fully implemented within the SVS system. The personalized data can then be loaded into the user's own SVS system, where it will re-create the same acoustical environment over a standard pair of stereo headphones. Since all the information is carried in the PRIR data, theoretically the user need never enter the measured sound room again.

For example, a music producer could listen to a mix away from the studio, confident that the headphone rendition is identical to that heard by the mixer. Audio post-production for film could also be made more efficient by, for example, enabling sound editors to use a virtualized dubbing stage to assess pre-mixes at audio work stations. Indeed mixes could be assessed in any number of environments at the push of a button. Moreover due to the inherent acoustic isolation of headphones, numerous sound editors, each using SVS, could work in the same physical location without disturbing each other.

The SVS system comprises three main parts for generating virtualised loudspeakers: a binaural measurement system, a streamlined multi-channel convolution engine and a simple, low-cost head-tracking system.

Personalised binaural room impulse (PRIR) measurement system

Headphone virtualisation systems operate on binaural room impulse response data. In the case of SVS each

binaural impulse records the acoustic signature of a single loudspeaker in the listening environment. However if the binaural impulse is generic, recorded for example using a dummy-head, then the subsequent accuracy of the virtualised speaker varies from listener to listener, and in no single case will be entirely accurate. On the other hand experimental evidence strongly suggests that virtual speakers created from individual binaural measurements cannot be distinguished, by the subject, from real speakers in the same environment. [4][5]

SVS has integrated test signals and simple measurement procedures to allow individual users to acquire their own personalised binaural data. However some binaural impulses from a dummy-head are provided as factory defaults, in particular for unconventional loudspeaker positions. Additional dummy head binaural data can also be measured using the SVS system.

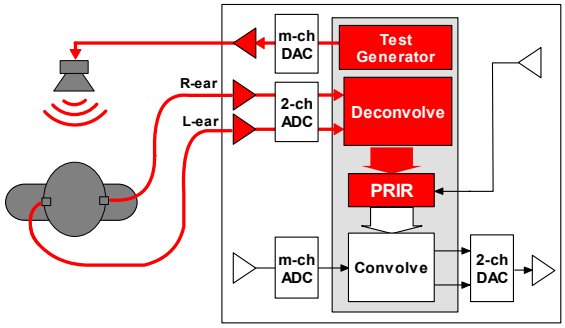

Figure 1. Smyth SVS has Integrated test signals and procedures for acquiring personal binaural room impulse response (PRIR) measurements.

PRIR look-angles

SVS simplifies the personalisation process by acquiring a sparse set of PRIR measurements for each active loudspeaker. Typically the system measures these responses for three different head positions, at approximately -30^0, 0^0 and $+30^0$ azimuthal angle. The head positions taken up during the measurements (herein referred to as look-angles) are often, for practical reasons, coincident with the front left, centre and right loudspeakers, and in this case allow accurate head-tracking around the critical central monitoring position. If the look-angles do coincide with the actual loudspeakers SVS can also generate auditory cues to indicate correct alignment of the subject's head with the three speakers during the acquisition. However, other look-angles can also be used if desired.

The three positions chosen allow simple rotational head-tracking to be accomplished by interpolation between the binaural data sets from each head position – for example, within the scope of the left and right speakers. This is a necessary but reasonable compromise. For critical listening the only viable monitoring position is looking straight ahead at the centre speaker, and thus is accurately virtualized using the SVS methodology. Head-tracking induced interpolation is only engaged when the user's head moves off centre.

The personalised binaural data can be stored on, and reloaded from, permanent memory devices, both within the SVS system itself and externally on SD cards. The data can therefore be moved easily between different units, for example recorded in a high quality reference listening room for use elsewhere.

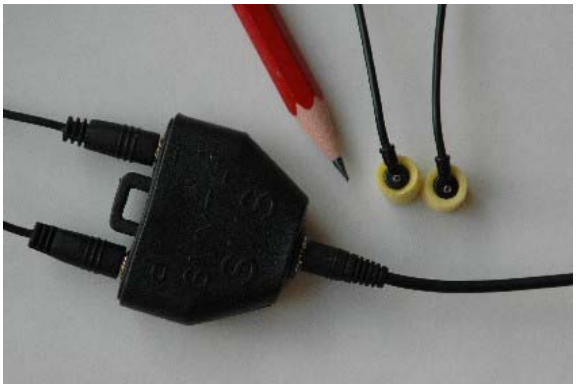

Figure 2 Custom SVS omni-directional microphones mounted in modified E.A.R. foam earplugs, used for measuring personal binaural room impulse responses (PRIR data).

PRIR measurement apparatus and procedure

Custom miniature omni-directional microphones, designed specifically for the SVS measurement procedures, are placed at the entrance to the left and right ear canals, in a blocked meatus configuration. With the head held stationary, room excitation signals (swept sine wave or MLS signals) are reproduced by each active speaker in turn, at a level that aims to maximise the SNR at the microphones. The binaural signals generated by the microphones are recorded back into the SVS processor and binaural room impulse responses for each speaker are calculated and stored. To further improve the measured SNR the excitation signal can be repeated multiple times and the PRIR data averaged.

This procedure is repeated for all three look-angles. The final data set consists of binaural room impulse responses, with the inter-aural time delay removed, for a maximum of eight active loudspeakers at three head orientations. The total measurement time for a 5.1ch speaker arrangement would normally be less than five minutes but may be varied by the user depending on the

ambient noise levels in the test room and the level of fidelity demanded by the application. Finally the azimuthal angle and elevation angle (if not equal to 0^0) of each active speaker is input by the user and stored as a single file with the measured PRIR data.

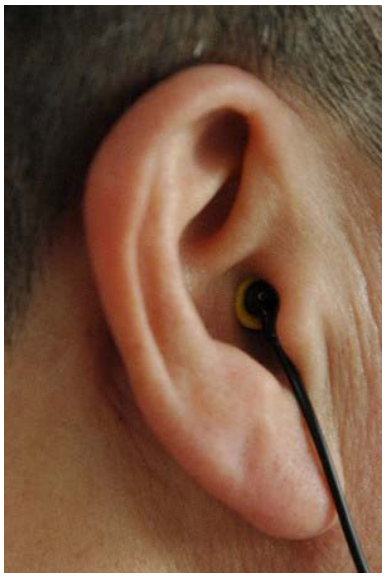

Figure 3 Custom SVS omni-directional microphone inserted in the entrance to the subject's ear canal.

Headphone EQ

Lastly, using the same microphones in the same in-ear location, the headphone-to-pinna response is recorded and a headphone equalisation filter for an individual user is generated and stored. Headphone equalisation filters may be needed to compensate for the filtering effects of the pinna when the virtualised audio is finally presented to the listener over headphones. The degree of equalisation required will ultimately depend on the design of the headphones. Circumaural headphones will need quite aggressive equalisation, compared to in-ear type headphones that should require no pinna-related compensation. In addition, improved headphone equalisation techniques specifically for loudspeaker virtualisation is the subject of on-going research.

To facilitate future developments in this important area, the SVS system can download its raw headphone-to-pinna response data and upload externally generated headphone equalisation filters, as an alternative to those generated internally.

Alternative measurement procedures

Similarly, the binaural data acquisition methodology integrated within SVS is not intrinsically linked to the core convolution engine, so that alternative, improved procedures and equipment can readily be adopted [6]. Improved accuracy in the measured binaural data translates directly to more accurate virtualisation. Provided the data is formatted correctly for uploading from the SD card, other methods of generating the binaural data, including modeling, are entirely possible.

DSP CONVOLUTION ENGINE

The actual real-time virtualization requires that the live audio signals, intended for loudspeaker reproduction, are filtered, or convolved, with the PRIR responses determined for each of these speakers.

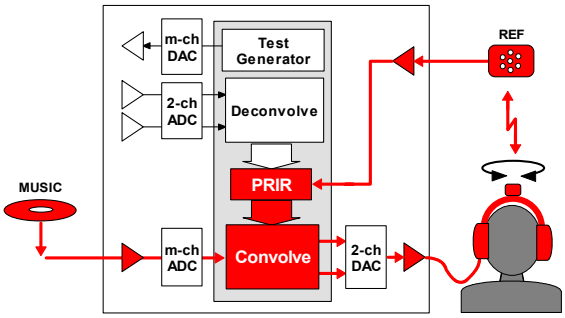

Figure 4. Convolution of PRIR data, either measured or up-loaded from external memory, with up to eight channels of incoming speaker source audio, generating a binaural headphone signal. PRIR interpolation and ITD insertion is controlled by a headphone-mounted optical head-tracker.

PRIR interpolation

Each measured PRIR contains all the information necessary to virtualise a single loudspeaker for one particular head orientation. The thee measured sets of PRIRs will virtualise all the active speakers in three head positions, looking left, right and centre. This allows dynamic head-tracking over the left-to-right speaker range, driven by the measured head azimuthal angle, by interpolating a new set of PRIRs between the three measured data sets.

ITD re-generation and insertion

After interpolation, the inter-aural time delay (ITD) is calculated for each binaural PRIR pair, driven also by the measured head angle. The calculated time delay is re-inserted back into each PRIR pair, at a sub-sample resolution, prior to convolution with the audio signal.

The two processes driven by the head-tracker, PRIR interpolation and ITD insertion, are highly dynamic, since head angles must be able to change slowly or quickly with no audible effect. However, whilst rapid head movements are quite common, in reality they impair the ability to listen critically. Therefore, provided no unnatural artifacts are audible, there is no need to

maintain a perfect acoustic sound field during rapid head movements. Nonetheless, the SVS system can track head movement of up to 360^0 per second without introducing audible artifacts.

After ITD insertion each binaural pair is convolved with an individual speaker signal, and the convolved signals, to a maximum of eight, are summed to a single pair of binaural signals. Before being output, the signals can also be optionally equalised, using the left-ear and right-ear filters calculated from the measured headphone-pinna impulse responses.

The final output is a personalised binaural signal suitable for presentation over headphones, containing up to eight virtual speakers. The position of all the virtual speakers are anchored to their true, external position, by tracking the orientation of the listener's head and dynamically changing the virtualised sound field. Using a tilt detector in the head-tracker it is possible to directly compare the real sound field (headphone off the head) with the virtual sound field (headphone on the head), at any head orientation between the left and right look-angles.

OPTICAL HEADTRACKER

Headphone based virtualisation, even using individualised binaural data, can easily cause confusion, particularly front-back reversals, if the virtual speakers rotate in lock-step with the subject's head. Head-tracking attempts to remove this confusion by continuously tracking and measuring the rotational angle of the user's head. This angle is used to calculate an interpolated PRIR data set appropriate for this new head orientation, and also adjusts the inter-aural time delay (ITD) between each of the binaural signal pairs for each of the active speakers. Finally the binaural signals are convolved with the source signals and the outputs summed for headphone presentation

Smyth Head-tracker is a relatively simple but effective optical system that determines the rotational (azimuth or yaw) angle of a detector situated on the listener's headphones, with respect to a fixed, infra-red (IR) reference transmitter. The IR transmitter defines an external reference plane and would normally sit on top of the centre speaker, TV or video screen. The head angle is calculated every 9ms, and this data is transmitted optically back to the reference transmitter and hence to the SVS system. The detector can be mounted on any set of headphones using an appropriate adaptor.

The range of operation is somewhat dependent on the background light conditions, particularly sunlight, and on the power output of the IR transmitter unit. Under normal video viewing conditions the standard (set-top) reference unit has an operational range of up to six metres. This can be extended to much larger distances using an upgraded IR transmitter unit. Head angles can be resolved to approximately 0.25^0 over a range of +/- 60^0 around the 0^0 reference, beyond which the detector cannot receive sufficient light from the transmitter. If the detector falls out of range, or gross data transmission errors occur, SVS reverts back to operating at a head angle of 0^0.

Figure 5 Smyth Optical Head-tracker mounted on a pair of entry-level Stax headphones, and the standard (set-top) reference IR transmitter device mounted on a laptop computer.

Personalisation: is it really necessary?

Personalised binaural data is acquired in an effort to guarantee a high level of performance for all users. Generic, non-individualised data could also be used, for example from a dummy head or from a subject who has good localisation abilities, but the acoustic accuracy of the virtual environment would then vary from user to user, depending on how similar their PRIRs were to the generic reference. This variability in the performance of the virtualisation system is reduced if personalised BRIR data is used, and the absolute level of performance for each user is also improved. [2]

However in some situations generic data may be useful. For example, if no personalised data exists for a particular speaker position, then dummy-head BRIR data may be added into a pre-existing personalised data set. The non-individualised virtual speaker may not sonically match the personalised speakers, and may not localise so accurately, but nonetheless would provide some level of performance. Mixing and matching between generic and personalised speakers adds flexibility to the system and allows for a greater degree of experimentation.

PRIR data: are three positions enough?

The three positional PRIR data sets, typically used by the SVS system, allow restricted head movements around the central monitoring position, sufficient to maintain the authenticity of the virtualisation. Nevertheless, interpolating between PRIRs does introduce some degree of inaccuracy. However experimental evidence [7] has shown that interpolation between two individualised HRTFs with an azimuthal separation of up to 30^0 does not introduce perceptible errors. Where SVS is used to virtualise 5.1ch loudspeaker arrangements, the PRIR separation is typically 30^0.

It should also be noted that any inaccuracy introduced is mitigated by two factors. First, the normal monitoring position is looking straight at the centre speaker, and here the interpolation distance is negligible. Therefore the PRIR data used for virtualisation during critical listening is almost identical to the measured data. Taking this a step further, the user can opt to temporarily disable the head-tracking, thereby completely removing inaccuracies introduced by PRIR interpolation.

ASSESSMENT OF VIRTUALISATION

Assessing the absolute performance of a headphone based loudspeaker virtualisation system is difficult. [8] Normal objective practice would dictate that a comparison be made to real loudspeakers, and that this reference be hidden from the assessor. However this concept is impractical since the headphones must be removed to hear the loudspeakers without distortion. This implies that listener bias, both negative and positive, cannot be excluded, and essentially equalises all opinion. Furthermore the lack of a hidden reference means that the expertise of the listener cannot easily be judged. Hence all listeners become experts.

This is also true in a different sense, in that virtualisation is wholly subjective and may be based on measured data that is unique to each user. Only the individual can assess the absolute accuracy of their own PRIR data. Non-individualised BRIR data, for example from a dummy head, only exacerbates this problem, since the accuracy of the virtualisation merely measures how closely each individual's BRIR data resembles that of the dummy.

Notwithstanding these very real issues, comparison to real loudspeakers is still demanded by professional users. SVS facilitates simple A/B comparisons, utilizing a tilt detector in the head-tracker to determine if the headphones are on, or off, the user's head. This can switch the unit into speaker pass-thru mode, muting the headphones, and vice-versa. Real and virtual speakers can thus be readily compared, individually or collectively, for any head orientation between the two outer look-angles.

VIRTUALISING THE FRONT CENTRE CHANNEL

Reproducing sound accurately from the centre front location should theoretically be the most difficult challenge for a headphone virtualisation system. Both aural and visual acuity is heightened in this region but, due to symmetry, only the unique pinna-related cues can be used to localise the sound. Therefore any errors in the virtualisation system, particularly with respect to HRTFs, tend to be readily distinguishable here when compared to reality. Since this is also the most significant source location for studio monitoring applications, evaluating the quality of the virtual centre speaker (and also the virtual phantom-stereo image) should claim priority. Some of the potential inaccuracies are listed below. These problems are not unique to the centre channel, merely easier to perceive there. Imaging should be assessed under both static and dynamic conditions.

- image
  - more diffuse than original
  - closer to the listener than original
  - inaccurate azimuthal displacement
  - more elevated than the original
  - tonally imbalanced

VIRTUALISATION PROBLEMS

Conflicting aural and visual cues

Even if headphone virtualisation is acoustically accurate, it can still cause confusion if the aural and visual impressions conflict. [8] If the likely source of a sound cannot be identified visually, it may be perceived as originating from behind the listener, irrespective of auditory cues to the contrary. Dynamic head-tracking strengthens the auditory cues considerably, but may not fully resolve the confusion, particularly if sounds appear to originate in free space. Simple visible markers, such as paper speakers placed at the apparent source positions, can help to resolve the remaining audio-visual perceptual conflicts. Generally the problems associated with conflicting cues become less important as users learn to trust their ears.

Lack of tactile bass response

Low frequency bass cannot be felt through the body with a headphone based virtual speaker system. This is somewhat unnatural for many experienced mixers. One solution is to use vibration plates under the chair, driven from the summed bass output of the real speaker audio signals.

One such system, demonstrated successfully with SVS, is Crowson Technology's TES-100 system that responds

to a range of low-frequency signals and thereby attempts to re-create many of the effects of bodily conducted sound. Even though it may not be entirely realistic, it has the ability to produce a sufficiently accurate level of tactile feedback at the correct frequencies for the sensation to be readily understood and used by mixers for timing purposes.

BEYOND 7.1 CH

For applications that require more than eight loudspeakers, for example live theatrical sound, multiple SVS systems can operate simultaneously for a single listener. Each system virtualises a different set of speakers, and the head-tracker data from a single user is daisy-chained through all the units. The multiple 2-channel line-level outputs from each system must be mixed externally to a single 2-channel headphone output. With analogue inputs all the SVS units will operate synchronised to within a single audio sampling period. This simple technique allows users to experiment with a large number of virtual speakers distributed in non-standard positions.

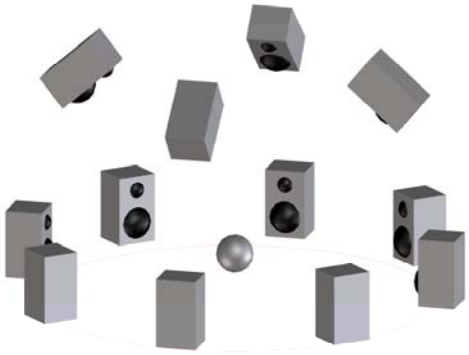

Figure 6 Example 12-ch loudspeaker arrangement suitable for virtualising with two SVS processors operating simultaneously.

CONCLUSION

Smyth SVS defines a complete system for virtualising loudspeakers over standard stereo headphones. Personalised binaural data is captured in the chosen environment and convolved with individual speaker channels. Dynamic aural cues, based on the measured azimuth angle of the headphones, are used to anchor the virtual sound field, strengthening the illusion of real speakers. The design goal for SVS was to replicate both the sound of loudspeakers and the real-world experience of listening to speakers, and the authors believe that in most practical respects this has been attained. Nonetheless, the performance of personalized systems such as SVS, and its suitability for particular monitoring applications, ultimately can only be judged by each individual listener.

REFERENCES

[1] Hartmann, W.M. and A. Wittenberg. 1996. "On the externalization of sound images" Journal of the Acoustical Society of America 99 (6): 3678–3688.

[2] Gardner, W.G., 2004. "Spatial Audio Reproduction: Toward Individualized Binaural Sound", The Bridge Vol. 34 No. 4: 37-42

[3] Horbach, U., Karamustafaoglu, A., Pellegrini, R., Mackensen, P., Theile, G., 1999. "Design and Applications of a Data-based Auralisation System for Surround Sound", 106th AES Convention, Munich 1999: preprint 4976

[4] Wightman, F.L. and Kistler, D.J. 1989a. "Headphone simulation of free-field listening I: stimulus synthesis". Journal of the Acoustical Society of America, 85(2)

[5] Wightman, F.L. and Kistler, D.J. 1989b. "Headphone simulation of free-field listening II: psychophysical validation". Journal of the Acoustical Society of America, 85(2)

[6] Farina A. 2000. "Simultaneous Measurement of Impulse Response and Distortion with a Swept-sine technique", 108th AES Convention, Paris, 2000: Preprint 5093

[7] Martin, R. and McAnally, K. 2007. "Interpolation of Head-Related Transfer Functions", Australian Government, DSTO-RR-0323

[8] Martens W.L., 2003. "Perceptual evaluation of filters controlling source direction: Customized and generalized HRTFs for binaural synthesis" Acoust. Sci. & Tech. 24, 5 (2003)

Figure 7. Smyth SVS Realiser, an 8-channel real-time headphone virtualisation processor. Shown with Smyth Optical Head-tracker mounted on Stax headphones, and the reference set-top IR transmitter.