Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Overview and comparative results of speech-based excitation signals for virtual localization and real-life applications

György Wersényi

Széchenyi István University, H-9026, Győr, Egyetem tér 1, Hungary

PACS: 43.66.Qp, 43.66.Pn

ABSTRACT

In virtual audio synthesis we use different excitation signals for listening tests. These tests are executed using headphone playback and real-time HRTF synthesis. Besides noise signals, speech is often used for tests especially for real life applications, such as mobile phones, various voice transmission lines, computer environments or accessories for the visually impaired, where speech intelligibility is important. This paper presents a summary of results of a listening test using different signals including speech in a virtual audio environment aimed at blind persons. Results show how speech contributes to the accessibility to computers and performs in a comparative test for virtual audio simulation. Furthermore, additional speech test signals are overviewed and introduced, such as speech-chorus signals, segmented spondees and the newly developed spearcons.

INTRODUCTION

The term “virtual audio” is used for comprehensive simulation of the sound field through headphones. In a wider sense, every playback system that includes headphone playback and where listeners have to orientate and act based on audio information only is regarded as a virtual audio display (VAD). Localization of the sound sources in such environments is the main challenge. Based on our objectives we use different excitation signals and playback methods. This paper focuses on the excitation signals, first of all, on speech. Speech samples can be used for various types of listening tests, and the main objective of the test determines what we can do and what we can not.

One of the main approaches of virtual audio synthesis is to create the most proper environment that man can simulate [1-3]. In this case, the spatial distribution of the virtual sound sources is crucial, directional information has to be transmitted properly and all localization errors that can appear have to be minimized. Well-known errors such as in-the-head localization and front-back reversals indicate problems with headphone playback itself [4-7]. It was shown that body movements, turning of the head or using head-tracking devices can help reducing these phenomena [8-11]. For the best simulation we use the filtering of the human Head-Related Transfer Functions (HRTFs) that can be measured on a manikin or on real human heads [12]. For the stimulus is broadband noise the most common choice, followed by narrow-band noise and speech.

Our current investigation is part of the GUIB project (Graphical User Interface for Blind Persons) [13-15]. On the one hand, localization behavior is tested in order to reduce in-the-head localization and front-back reversal rates to enhance the virtual audio environment that blind users use during computer usage. On the other hand, various sound samples are selected and tested in order to extend their special software applications. Both investigations include speech as the most important sound sample that helps blind users to access personal computers. Speech is the basis for text-to-speech (TTS) programs and screen-readers that offer vocal feedback of the visual screen. The most popular application is JAWS (Job Access With Speech) offering multilingual speech and some sound samples as well [16].

This paper first presents some results from a listening test. To decrease error rates we emulated small, unintentional head-movements of the listener’s head by changing and updating the HRTFs only. This method does not need any additional hardware, head-tracking sensors or similar. Speech is applied and a comparison can be made with noise excitation and MS Windows warning signals. The second part of the paper introduces additional methods and speech samples for menu navigation and for orientation in a virtual environment by focusing on the needs of the blind community.

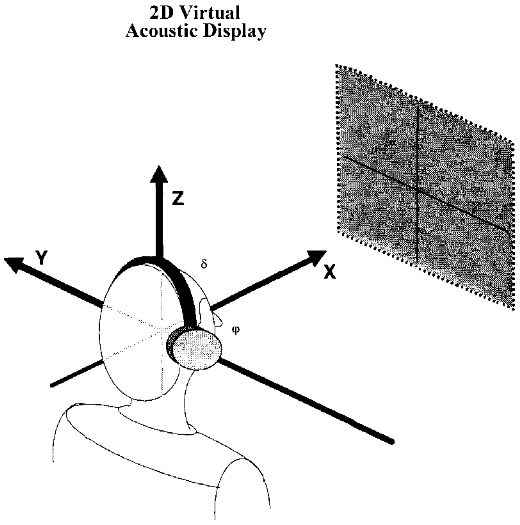
EMULATION OF HEAD-TRACKING FOR REDUCING HEADPHONE ERROR RATES

In our experiment seven different types of excitation signals were tested including broadband noise, filtered noise, speech and well-known sounds of MS Windows [17, 18]. Fifty untrained subjects participated, all with normal hearing and vision (the investigation involving the blinds is still in progress). Signals were presented by the BEACHTRON sound card using real-time filtering of HRTFs from a “good localizer" and the Sennheiser HD540 headphone (Fig.1.) [13, 14]. Subjects were asked to report about in-the-head localization and front-back reversals. The sound source was simulated first stationary in front of the listener, and then randomly moved by ±2 degrees at three different presentation speeds (fast, slow and average). Possible sound source locations during emulation are represented by red dots (Fig. 2). The virtual sound source is randomly moved from one location to another at various speeds emulating a similar movement of the listener's head in real life. Instead of moving the head (which is not helpful in a virtual environment at all), the sound source location is updated via the applied HRTFs. With this method we may reduce error rates without any additional hardware or head-tracking devices, sensors. The speech signal was a female voice sample saying a meaningful word ("welcome").

Results

Tables 1 and 2 show results for in-the-head localization and front–back errors for all stimuli and presentation speeds, respectively [18]. The numbers indicate correct answers during the emulation for subjects who gave incorrect answers without emulation. Summarizing all the results, we concluded that for about 22% of the subjects this kind of emulation was helpful for reducing in-the-head localization, and for about 14-16% for solving front-back discrimination. Presentation speed (the speed of the movement of the emulated sound source) is also important. Speech as an excitation signal is less effective than noise or even warning signals. For speech, average or slow presentation is recommended. While speech is externalized more frequently, the simulation of head-movement has a bigger impact on Windows-sounds. There is no clear evidence that familiar sounds from the operating system would be better than speech or noise. They may be more informative in meaning, but localization performance depends mainly on spectral content and second, on presentation speed. Basically, the same conclusions can be drawn for front–back rates. Increase due to the emulation can be observed for the Windows-sounds at medium speed and for speech at fast speed. It is common that by using slow presentation speed, the emulation is more helpful.

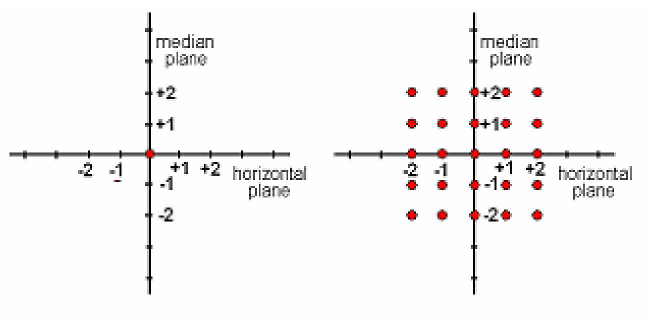
.



Source: (Crispien, 1993)

Figure 1. Illustration of the virtual audio environment. The acoustic surface is parallel to the Z–Y-plane. The origin (the reference location) is in the front of the listener.

.



Source: (Wersényi, 2007, 2009)

Figure 2. Illustration of the emulation. A stationary sound source simulated in front of the listener (left). Possible source locations in case of an emulated sound source movement of ±2 degrees in front of the listener (right).

Table 1. Results for in-the-head localization of 50 subjects for all stimuli and presentation speed. WN indicates white noise, W1 is "exclamation.wav", W2 is "critical stop.wav" and W3 is the "recycle bin.wav" of Windows XP.

IHL	WN	1500 Hz LPF	7000 Hz HPF	Speech	W1	W2	W3
Fast	14	10	8	**11**	8	16	10
Avg	17	14	9	**7**	13	8	9
Slow	17	9	11	**9**	8	11	11

Source: (Wersényi, 2009)

Table 2. Results for front–back reversals of 50 subjects for all stimuli and presentation speed. WN indicates white noise, W1 is "exclamation.wav", W2 is "critical stop.wav" and W3 is the "recycle bin.wav" of Windows XP.

IHL	WN	1500 Hz LPF	7000 Hz HPF	Speech	W1	W2	W3
Fast	2	7	7	**4**	8	7	7
Avg	9	9	5	**1**	10	9	9
Slow	6	11	9	**11**	7	6	7

Source: (Wersényi, 2009)

ADDITIONAL SPEECH SAMPLES FOR LISTENING TESTS

Our experiment above supported the observation that for virtual localization performance is spectral content more important than "meaning". Speech is not the best signal for optimal localization (because of the smaller bandwidth), but for applications that use mainly speech signals it is recommended to measure localization performance and error rates. This section presents additional speech signals and design criteria considering the needs of the particular application.

Averaged speech noise

Averaged speech samples are very popular for the determination of transmission properties, such as transfer functions or distortion measurements. Using speech databases (e.g. BABEL) we created averaged speech samples using the speech chorus method for Hungarian, English and German language already [19]. These samples contain male and female speakers of 15 speakers each, to create a speech noise signal (also called speech-chorus method). Spectral properties of different languages, gender, and age can be tested by analyzing them in the frequency domain. Furthermore, transmission properties of different speech transmission lines can be tested fast and reliable.

Spondees

Spondees can be designed for every language based on speech databases as well. In this case, words are used and cutted in different ways to simulate transmission loss by creating „missing speech” parts inside a spondeus. Restoration by the human recognition can be tested by filling these gaps with different noise signals, e.g. white noise, speech noise, or other [20-23].

Table 3 shows some spondees created for English [20]. The spondees were not created based on any statistical facts of English language. The spondees on list A are just possibilities that we could think of (each syllable being a meaningful word on its own, but without regard to whether the syllable was CVC, CV, or even VC), though we included only the words with “long” vowels, not short vowels. The "C" refers to both individual consonants and consonant blends - which are very common in the English language. The spondees on lists B and C have the same first syllable (on list B) or the same second syllable (list C) as in list A, while the other syllable in each pair has the same vowel as the original syllable (from list A) in either the second (for list B) or the first (for list C) position. List D has no match for either the first or second position, and the vowels intentionally do not match the vowels in either the first or second position in list A. The lists are constructed thusly:

A: A collection of real English spondee words

B: The initial consonant-vowel pair matches list A

C: The final vowel-consonant pair matches list A

D: Both the initial CV and the final VC match list A.

These spondees were then recorded (400 all together) by two male speakers. From the 400 recordings from each speaker, 16 lists of 25 spondees were created. In an American investigation the spondees had their centers cut out from mid-vowel to mid-vowel. In this case, mid-vowel means the geometric center of the vowel. All spondees were cut again, leaving 25% of each vowel, and again leaving 75% of each vowel. These blank-centered spondees are tests in their own right, but three different kinds of fillers were made. One filler had speech-spectrum noise with a flat envelope. Another filler had speech-spectrum noise with an envelope which matched the speech envelope of the spondee. A final filler matched the fundamental frequency of the speaker’s voice, with a simple sawtooth wave. A similar investigation is being made using Hungarian spondees.

Preliminary results for Hungarian language showed that fillers are not helpful at all; they rather are confusing and misleading. Restoration of the missing consonants depends strongly of the cutting percentage. The end consonant of the first syllable (that is the consonant following the first vowel) is less predictable than the first consonant of the second syllable (by 25% cut).

Table 3. Examples for English spondees for listening tests. Words were cutted in the vowels, at 25%, 50% or 75%.

LIST A: SPONDEES		**LIST B: SAME FIRST**	
AIR	PLANE	AIR	MATE
AIR	PLANE	AIR	MATE
ARM	CHAIR	ARM	HAIRS
BACK	PACK	BACK	TAG
BALL	PARK	BALL	YARD
BAND	STAND	BAND	PATCH
BANK	ROLL	BANK	COLT
BASE	BALL	BASE	POND
BAT	BOY	BAT	COIN
BATH	ROBE	BATH	FLOAT
BATH	MAT	BATH	LAND
BIRD	BATH	BIRD	VAN
BIRTH	DAY	BIRTH	VASE
BLACK	BOARD	BLACK	FORT
CAKE	WALK	CAKE	DOLL
CORK	SCREW	CORK	BOOTH
LIST C: SAME SECOND		**LIST D: SAME START/ END**	
BEARS	PLANE	AID	RAIN
YARD	CHAIR	ART	FAIR
SAND	PACK	BATH	STACK
DOG	PARK	BALD	LARK
MATH	STAND	BANK	HAND
SHADE	ROLL	BAND	HOLE
TAIL	BALL	BAIT	FALL
PACK	BOY	BANK	TOY
FAN	ROBE	BAND	GLOBE
RAG	MAT	BAND	PAT
WORK	BATH	BIRTH	PATH
WORM	DAY	BIRD	STAY
LAND	BOARD	BLANK	CORD
CHAIN	WALK	CAPE	LOCK
FORT	SCREW	CORN	DEW

Spearcons

Blind users resent that their text-to-speech application often speaks slow, some words corresponding to icons or events are not really mapped in the meaning and they could speed up their workflow with better sounds and/or speech samples. Earcons and auditory icons are sound samples representing visual icons or events on the screen [24-25]. They are frequently used to represent information of a computer screen, e.g. the sound linked to „empty recycle bin" or to „critical stop". Similarly, spearcons were designed as "speech-based earcons" [26-28]. A MATLAB routine was created to compress speech samples. English spearcons were successfully tested in Nokia cellular phones, as long Hungarian spearcons are being tested by the blind community in order to extend the computer environment they use [29, 30]. Spearcons are not just „speeded up" speech. Intelligibility is not a requirement in this case, users can get used to how the spearcons sounds like, even without understanding them. Although, spearcons are language dependent, they are easy to create (record manually and compressed with MATLAB) and they can be used next to the earcons in applications where extension of the sound database is required. Table 4 shows some spearcons and compress ratio.

Table 4. List of services and features for Hungarian spearcons introduced to blind users. The length and compress ratio is also shown. Original recording was made by a male speaker in 16 bit, 44100 Hz resolution using a Sennheiser ME62 microphone [29, 30].

Spearcon	Duration (original) [sec]	Duration (compressed) [sec]	Compress ratio [%]
Close	0,87	0,302	65,3
Open	0,812	0,288	64,5
Save	0,687	0,257	62,6
Save as	1,125	0,362	67,8
Search	0,694	0,258	62,8
Copy	0,818	0,289	64,7
Move	0,748	0,272	63,6
Delete	0,661	0,25	62,2
Print	0,752	0,273	63,7
Download	0,853	0,298	65
Stop	0,908	0,311	65,8
Word	0,576	0,228	60,4
Excel	0,599	0,234	60,9
Database	1	0,333	66,7
Start Menu	0,734	0,268	63,5
Browser	0,845	0,296	65
E-Mail	0,545	0,22	59,6

Source: (Wersényi, 2008)

Our recent investigation using spearcons included the Hungarian blind community as well [30, 31]. Spearcons performed relatively well together with auditory icons we created for the most important icons and events on the screen. Furthermore, the auditory emoticons were introduced. These are non-speech human voice(s), sometimes extended and combined with other sounds in the background. They are related to the auditory icons the most, using human non-verbal voice samples with emotional load. Auditory emoticons – just like the visual emoticons - are language independent and they can be interpreted easily, such as the sound of laughter or crying can be used as an auditory emoticon.

Auditory icons, earcons, spearcons and auditory emoticons, or structured combination of environmental sounds, music, non-speech audio or even speech can create good iconic representations.

SUMMARY

This paper overviewed speech as a possible excitation signal for virtual audio synthesis. Different types of speech signals were presented that can be used for telephone applications, menu navigation, blind users' accessibility or transmission line measurements. Involving also the blind community, a simulation of head-tracking was tested using headphone playback. This emulation of head-tracking instead of using an external device seemed to be helpful for about 22% for reducing in-the-head localization and about 14-16% for reducing front-back reversals. Speech, however, is still not the best excitation signal in spatial simulation: white noise and well-known warning signals of Windows perform better in the listening tests where localization tasks have to be solved.

On the other hand, where intelligibility or accessibility enhancement is important, speech signals are widely used mostly without an accurate spatial simulation. Besides meaningful words and speech noise signals, spondees and spearcons are good candidates for listening tests and real-life applications. These language-dependent samples assist recognition- and restoration measurements as well as help to extend and enhance accessibility of computer usage.

REFERENCES

1 J. Blauert, *SpatialHearing* (MIT Press, MA, 1983)
2 P. Minnaar, J. Plogsties and F. Christensen, "Directional Resolution of Head-Related Transfer Functions Required in Binaural Synthesis" *J. Audio Eng. Soc.* **53(10)**, 919-929 (2005).
3 F. Chen, "Localization of 3-D Sound Presented through Headphone - Duration of Sound Presentation and Localization Accuracy" *J. Audio Eng. Soc.* **51(12)**, 1163-1171 (2003)
4 F.E. Toole, "In-head localization of acoustic images" *J. Acoust. Soc. Am.* **48**, 943-949 (1969)
5. N. Sakamoto, T. Gotoh and Y. Kimura,": On „out-of-head localization" in headphone listening" *J. Audio Eng. Soc.* **24**, 710-716 (1976)
6 P.A. Hill, P.A. Nelson and O. Kirkeby, "Resolution of front-back confusion in virtual acoustic imaging systems" *J. Acoust. Soc. Am.* **108(6)**, 2901-2910 (2000)
7 Gy. Wersényi, "On the improvement of virtual localization in vertical directions using HRTF synthesis and additional filtering" *Proceedings of the 19th International*

Congress on Acoustics (ICA), Madrid, Spain, 2007 /5 pages/

8 W. Noble, "Auditory localization in the vertical plane: Accuracy and constraint on bodily movement" *J. Acoust. Soc. Am.* **82**, 1631-1636 (1987)

9 P. Minnaar, S.K. Olesen, F. Christensen and H. Moller, "The importance of head movements for binaural room synthesis", *Proc. in. Conf. on Auditory Display*, Espoo Finland, 21-25 (2001)

10 D.R. Perrott, H. Ambarsoom and J. Tucker, "Changes in Head Position as a measure of auditory localization performance: auditory psychomotor coordination under monaural and binaural listening conditions" *J. Acoust. Soc. Am.* **82**, 1637-1645 (1987)

11 D.R. Begault, E. Wenzel and M. Anderson, "Direct Comparison of the Impact of Head Tracking Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source" *J. Audio Eng. Soc.* **49(10)**, 904-917 (2001)

12 M. Kleiner, B.I. Dalenbäck and P.Svensson, "Auralization – an overview" *J. Audio Eng. Soc.* **41**, 861-875 (1993)

13 K. Crispien and H. Petrie, "Providing Access to GUI's Using Multimedia System – Based on Spatial Audio Representation" *J. Audio Eng. Soc. 95th Convention Preprint*, New York, (1993)

14 Gy. Wersényi, "Localization in a HRTF-based Minimum Audible Angle Listening Test on a 2D Sound Screen for GUIB Applications" *Audio Engineering Society (AES) Convention Preprint Paper, Nr.5902*, Presented at the 115th Convention, New York, (2003)

15 Gy. Wersényi, "Localization in a HRTF-based Minimum-Audible-Angle Listening Test for GUIB Applications" *Electronic Journal of "Technical Acoustics" (EJTA)*, Russia, http://www.ejta.org, 2007, 1. /16 pages/

16 http://www.freedomscientific.com

17 Gy. Wersényi, "Simulation of small head-movements on a virtual audio display using headphone playback and HRTF synthesis" *Proceedings of the 13th International Conference on Auditory Display (ICAD 07)*, 73-78, Montréal, Canada, (2007)

18 Gy. Wersényi, "Effect of Emulated Head-Tracking for Reducing Localization Errors in Virtual Audio Simulation" *IEEE Transactions on Audio, Speech and Language Processing (ASLP)* **17(2)**, 247-252 (2009)

19 Gy. Wersényi and A. Illényi, "Averaged speech signal samples generated by speech chorus method" *Proceedings of the International Békésy Centenary Conference on hearing and related sciences*, 115-120, Budapest (1999)

20 P. Divenyi and A. Lammert, "Do we perceive articulatory gestures when we listen to speech?" *J. Acoust. Soc. Am.* **123**, 3179(A) (2008)

21 R.M. Warren, „Perceptual restoration of missing speech sounds" *Science* **167**, 392-393 (1970)

22 J.A. Bashford, K.R. Riener, and R.M. Warren, „Increasing the intelligibility of speech through multiple phonemic restorations" *Perception & Psychophysics* **51**, 211-217 (1992)

23 J.A. Bashford and R.M. Warren, "Perceptual synthesis of deleted phonemes" *In J.J. Wolf and D.H. Klatt (Eds.), Speech Communication Papers*. New York: Acoustical Society of America, 423-426 (1979)

24 D. Burger, C. Mazurier, S. Cesarano and J. Sagot, "The design of interactive auditory learning tools" *Non-visual Human-Computer Interactions* **228**, 97-114 (1993)

25 M.M. Blattner, D.A. Sumikawa and R.M. Greenberg, "Earcons and Icons: their structure and common design principles" *Human-Computer Interaction* **4(1)**, 11-44 (1989)

26 M.L.M. Vargas and S. Anderson, "Combining speech and earcons to assist menu navigation" *Proceedings of the 9th International Conference on Auditory Display (ICAD 03)*, Boston, USA (2003)

27 B.N. Walker, A. Nance and J. Lindsay, "Spearcons: Speech-based eracons improve navigation performance in auditory menus" *Proceedings of the 12th International Conference on Auditory Display (ICAD 06)*, 63-68, London, UK, (2006)

28 D.K. Palladino and B.N. Walker, "Learning rates for auditory menus enhanced with spearcons versus earcons" *Proceedings of the 13th International Conference on Auditory Display (ICAD 07)*, Montréal, Canada, (2007)

29 Gy. Wersényi, "Evaluation of user habits for creating auditory representations of different software applications for blind persons" *Proceedings of the 14th International Conference on Auditory Display (ICAD 08)*, Paris, France, (2008)

30 Gy. Wersényi, "Evaluation of auditory representations for selected applications of a Graphical User Interface" *Proceedings of the 15th International Conference on Auditory Display (ICAD 09)*, 41-48, Cobenhavn, Denmark (2009)

31 Gy. Wersényi, "Auditory Representations of a Graphical User Interface for a better Human-Computer Interaction" in CMMR/ICAD post proceedings edition, Lecture Notes in Computer Science vol.5954, Springer Verlag, 2010, 24 pages.