PRELIMINARY VALIDATION OF THE REVISED STI MALE FOR THE ENGLISH LANGUAGE

L Morales The Acoustics, Group, London South Bank University.
S Dance The Acoustics Group, London South Bank University.
G Leembruggen ICE Design, Acoustic Directions, University of Sydney

1. INTRODUCTION

Extensive work has been undertaken to validate the Speech Transmission Index by comparing the results with measured intelligibility scores. The validation of the relationship between the subjective and objective aspects the STI for the English language that are included in the current STI standard (IEC60268-16:2011[1]) was taken from Anderson and Kalb[2] in which the subjects listened to a monaural speech program through headphones.

In the Anderson and Kalb[2] validation, an average of female and male spectra was used in the STI calculations and therefore the STI result did not consider the male-specific octave band weightings or redundancy factors that are now utilised in STI calculations.

This paper presents a validation of the STI method with English phonetically-balanced (PB) word lists in which one-hundred subjects listened to speech material subject to natural reverberation within a real space. An additional experiment was also designed in which eighty subjects listened to recorded binaural speech program through headphones. The difference between the real-life intelligibility tests and those carried out with headphones was analysed.

Part of the motivation behind our study is to explore the apparent poor correlation in reverberant situations between the measured STI and subjective speech intelligibility, which has been reported in [3], [4], [5].

2. VALIDATION OF THE STI WITH REAL LIFE LISTENING

2.1 Previous validation of the STI for the English language.

Anderson and Kalb's[2] 1987 validation used Harvard phonetically balanced (PB) word lists. For the test, words were originally recorded with a single microphone by three male and two female speakers in a quiet test room. Reverberation and noise were added electronically to the original recordings to degrade the speech signal. Two sets of reverberation conditions were created with a reverberation unit. The reverberation unit was set for a delay of 95ms with a repeat of ten reverberations, the strength of the first reverberation being equal to the original signal and each succeeding reverberation being half of the previous signal. Band pass limiting was also used to degrade the speech signal.

The degraded word lists were presented to three male and three female listeners through headphones in a monaural format. The listeners were familiarized with the words so that under the best listening conditions they were consistently able to identify a minimum of 95% of the words. No significant differences were found between the intelligibility scores resulting from male or female talkers.

2.2 Method.

The real-life intelligibility tests in this study took place in a reverberant chamber of 202.7 m^3 in volume and dimensions of 7.6m x 6.35m x 4.2m (L x W x H).The STIs were measured using the Indirect Method[1] (Schroeder equation using the impulse response) with a B&K 2236 SLM equipped with a B&K 4188 microphone, fed to WinMLS 2007.

Five different reverberation scenarios were chosen to degrade the speech signal. The five scenarios were achieved by exposing built-in absorptive surfaces within the chamber and also by introducing different amounts of absorptive panels. The absorptive panels were located at the front and the back walls of the chamber and as far as possible from the listening positions. For each reverberation scenario, a selected number of listening positions were distributed throughout the floor of the chamber in order to achieve ten different STI scenarios ranging from 0.36 to 0.70 STI.

Figure 1. Reverberant chamber in one of the reverberation scenarios. A few of the listening positions and the monitor loudspeaker used for the listening tests are shown.

A high-quality active studio monitor loudspeaker was used for the tests. The monitor loudspeaker was placed 1.5m away from any reflective surface and its acoustic axis was located at approximately 1.2m from the floor. A total of five PB lists were recorded in anechoic conditions using five male native English speakers. The PB words were embedded in a carrier sentence of the form: “Write the word……, please”. The sentences were recorded with three-second gaps between them and pronounced without stress. The standard PB lists are included in ANSI S3.2-2009[6].

One hundred English students, sixty-two females and thirty-eight males between 18 and 30 years old participated in the real-life speech intelligibility tests. Prior to the tests, all the students undertook a standard Bekesy audiometry test and responded to a questionnaire related to their hearing. All students had normal hearing and most presented with excellent hearing. None of the listeners were trained for this listening test nor had they participated in similar tests before.

The listeners sat on chairs at one of the ten locations facing the monitor loudspeaker and were instructed to write down the words they heard. The recorded PB lists were played from a high quality CD player through the monitor loudspeaker under quiet conditions. To avoid noise contamination, the lists were played at an average level of 76dBA with a deviation of less than 1dBA at any of the listening positions. Each subject listened to one of the PB lists only. Ten listeners were used per listening position.

2.3 Results.

Figure 2 below shows the relationships between the PB-word score results and the monaural STI measured with the omni-directional microphone for the ten listening positions. The graphs also include the relationship between the STI and the Standard PB-word lists that is currently used in IEC60268-16:2011.

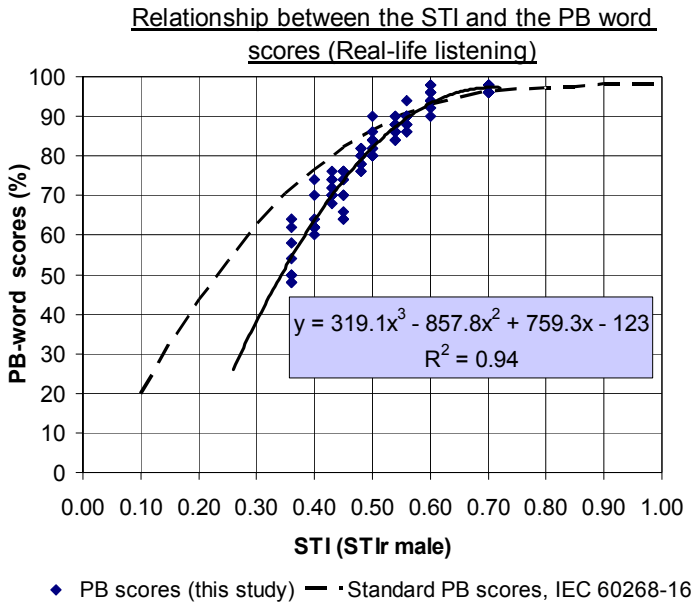

Figure 2a. Relationship between the monaural measured STI (STIr male) and the score results of the real life PB word intelligibility tests.

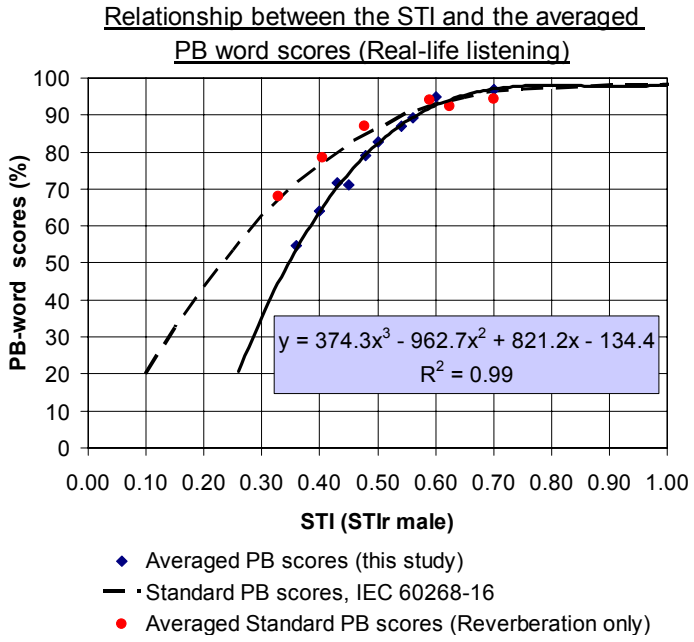

Figure 2b. Relationship between the monaural measured STI (STIr male) and the averaged score results of the real life PB word intelligibility tests. The averaged PB results for reverberation distortions taken from Anderson and Kalb[2] are also given.

As noted earlier, the Anderson and Kalb relationship between STI and PB-word scores was obtained using reverberation, noise distortion and band-pass limiting distortions, with the results for the 6 listeners being averaged for each scenario. As only degradation by reverberation was considered in our study, the STI/PB relationship for reverberation only conditions was extracted from the Anderson and Kalb data, and is shown in Figure 2b for comparison with our results.

The relationship between averaged PB-word scores and the STI that we found is given by the following polynomial:

$$PB\ scores = 374.3 \cdot STI^3 - 962.7 \cdot STI^2 + 821.2 \cdot STI - 134.4.$$

The overall standard deviation difference between the averaged scores and the regression curve for this study was calculated to be 1.4% of the scores with the coefficient of determination $R^2 = 0.99$.

The PB word score results were also compared with the measured binaural STI for each listening position. When performing binaural measurements, the suggested approach is to provide the STI value for the best ear[7] and the method of IEC60268-16:2011 Figure 3 below gives the relationship between the averaged PB results of this study and the best ear STI results selected from the binaural measurements. The averaged PB results for reverberation distortions taken from Anderson and Kalb are also included.

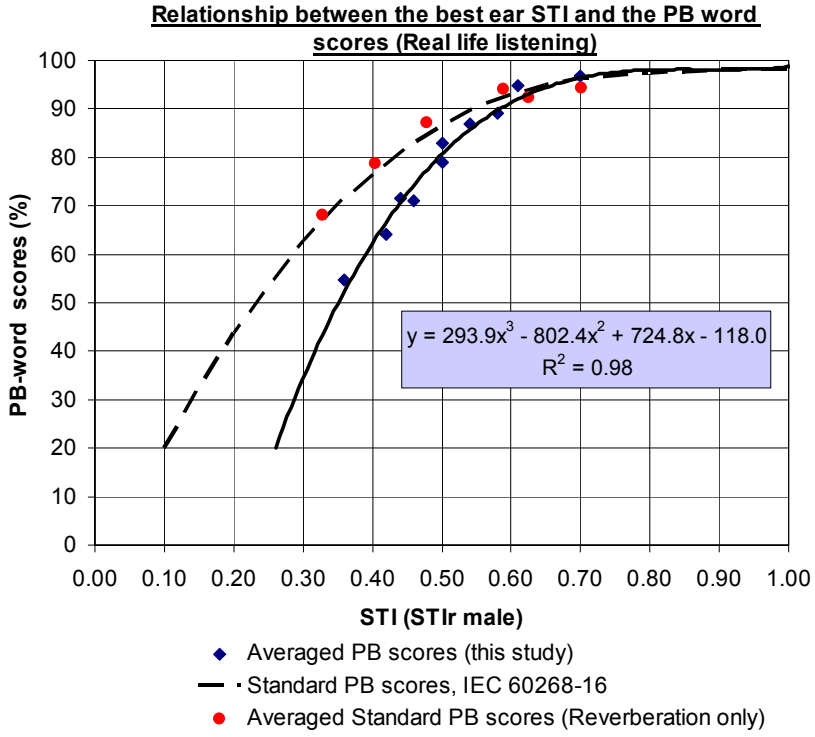

Figure 3. Relationship between the best-ear binaural measured STI and the averaged score results of the real-life PB word intelligibility tests.

The relationship between PB-word scores and the best ear binaural STI is given by the following polynomial:

$$PB\ scores = 293.9 \cdot STI^3 - 802.4 \cdot STI^2 + 724.8 \cdot STI - 118.0$$

The overall standard deviation difference between the averaged scores and the curve for this study was calculated to be 2.1% of the scores with the coefficient of determination $R^2 = 0.98$.

2.4 Discussion.

The results of the present study show lower averaged STI scores than those indicated by the standard PB curve with values below 0.60 STI. Above approximately 0.60 STI, the relationship between the averaged PB scores and STI is very similar. With values below 0.60 STI, the difference between the scores increases as the STI decreases, as seen in figures 2 and 3. Essentially, this indicates a lower level of subjective speech intelligibility for a given STI score. It is noteworthy that the even with the binaural STI measurement, the PB word scores are lower in our study than the Anderson and Kalb study.

Research has found female speech to be more intelligible than male speech[8]. Noting that the Standard PB/STI relationship was obtained by averaging the scores from male and female talkers, higher scores could be expected with the Standard PB method from the contribution of the female speech compared to this study which used male speech only. However, Anderson and Kalb reported that no significant differences were found between male and female talkers. We therefore conclude that the inclusion of female speech does not explain the higher PB scores of the Standard PB validation compared to this study.

In Anderson and Kalb's study, the listeners were trained until they could achieve a minimum of 95% of the scores under the best listening conditions. This training could have helped the listeners to guess the words under the most difficult conditions. Therefore, the lower PB scores found in this study could be a result of the lack of training of the participants, which is a more realistic representation of real-life situations.

3. VALIDATION OF THE STI WITH HEADPHONES

3.1 Method.

The PB lists used for the real-life listening tests were played through a monitor loudspeaker in the reverberant chamber and recorded binaurally at eight of the real-life listening positions. All the PB lists were replayed at the same levels as those used for the real-life listening tests.

Although the HATS' microphone is positioned at the entrance to the ear canal, the HATS' output is equalised by the device to introduce the frequency response and gain of the ear canal. To produce an equalisation to flatten the frequency response at the entrance of the ear canal of the listeners[9,10] the frequency response of the system, HATS and headphones, was measured using WinMLS 2007 and impulse response techniques. This process captured the combined effect of the HATS' ear canal gain on the PB recordings and the frequency response of the playback headphones in a single IR. The recorded PB word lists were pre-equalised with the inverse of this response.

Sennheiser HD-650 headphones were used for the listening tests, which are an open type, also called *free-air equivalent coupling* to the ear, and are known to provide a diffuse field at the listener' ears.

Taking into account concerns associated with the measurement of headphones[11], the headphones were re-positioned six times on the HATS ears and a new IR captured. Figure 4 shows the HATS and headphones used for measurements and their combined frequency responses for each ear. The responses were measured six times for each ear.

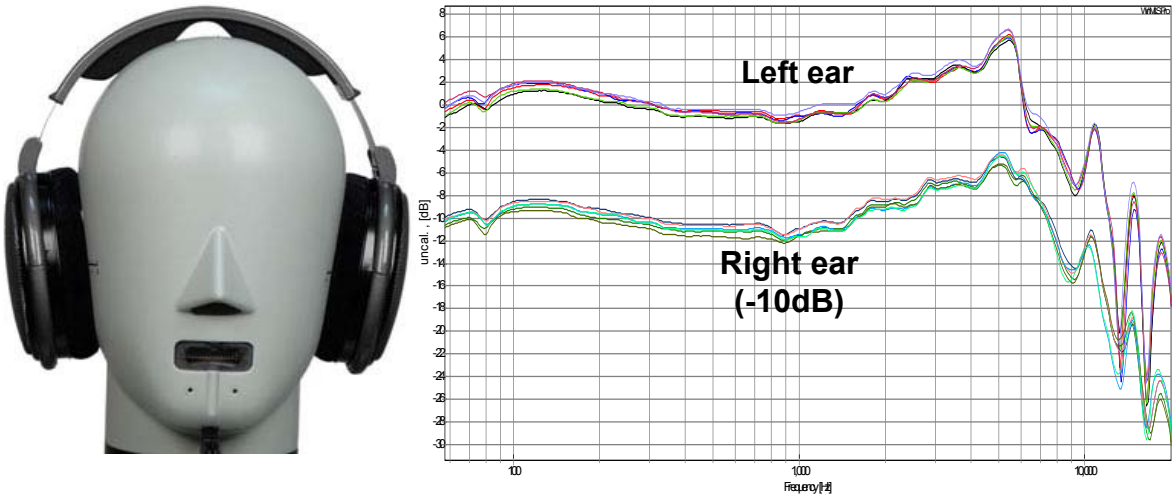

Figure 4. B&K 4100 HATS and Sennheiser HD-650 on the left and their combined frequency responses for both ears are shown on the right. The responses are presented un-normalized and without smoothing effects. The right ear responses are reduced by 10 dB for ease of viewing.

Given that HATS is designed to mimic the average listener physiology, this equalisation method would provide a flat frequency response to the entrance of the ear canal of an average listener. However, the pinna of listeners and the coupling between the headphones and their heads would be different from one listener to another. These differences would make individual equalisation preferable when correcting the frequency response of the headphones[10].

Due to the high number of listeners used for the intelligibility tests, practicality precluded individual equalisation. However, the error in the frequency response introduced with non-individual equalization has been found to be negligible for frequencies up to 1 kHz and less than ± 3.0dB for frequencies up to 5 kHz[12]. Only at high frequencies, 6-7 kHz, the errors due to non-individual equalization exceed ± 5.0dB [12].

Eighty native English students, forty-three females and thirty-seven males between 18 and 30 years old participated in the headphone speech intelligibility tests. Prior to the tests, all students undertook a standard Bekesy audiometry test and responded to a questionnaire related to their hearing. All students had normal hearing and most presented with excellent hearing. Similar to the real life listening tests, none of the listeners were trained for the listening test nor had they participated in similar tests before. Each of the 80 students listened to one of the PB lists only.

For the listening test, the headphones were connected to the stereo headphones output of a high quality CD player (Denon DCD-1500-AE). The frequency response of CD headphone output was measured for both channels and found to be flat within 0.5 dB over the range 50 Hz to 20 kHz. Additionally, the level balance between the left and the right channels of the headphone output was measured and was found to be within 0.2dB at all frequencies.

The speech level was adjusted so that the headphones produced levels within 1dBA as those presented on the real-life tests for all the scenarios. Similar to the real-life intelligibility tests the PB recordings presented to the listeners were contaminated with reverberation only. The intelligibility tests were conducted in a very quiet room where the listeners were instructed to write down the words they heard.

3.2 Results: Headphones versus real-life listening test

Figure 5 below compares the averaged results with headphones with the averaged results obtained for eight of the real-life listening scenarios at the same eight positions, both for monaural STI measurements. The scale of the horizontal and vertical axis (STI and percentage PB scores respectively), was restricted to the values obtained in this study in order to provide a better view of the results.

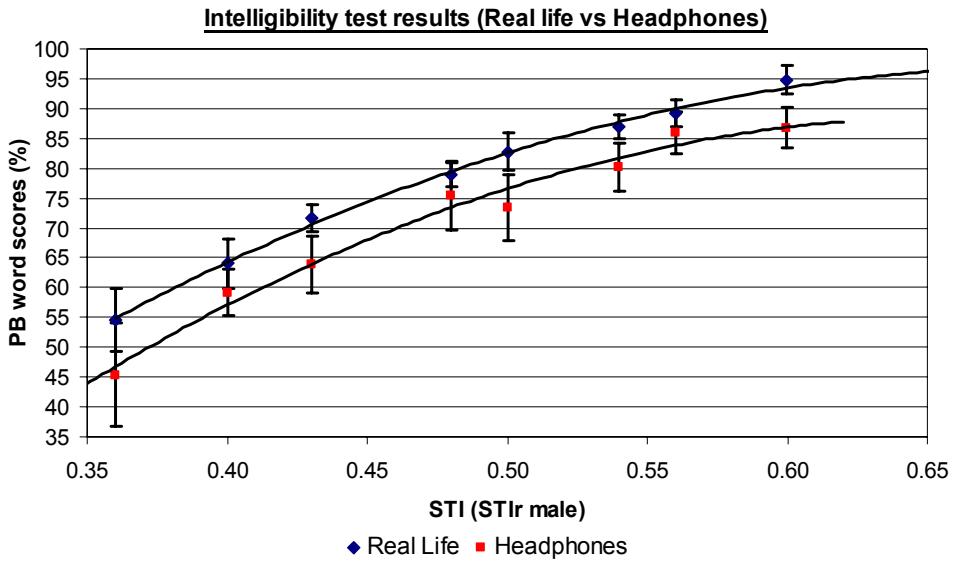

Figure 5. Averaged PB-word score results obtained with the real space and the headphone methods for the eight STI scenarios. The error bars indicate one standard deviation.

The overall standard deviation difference between the averaged scores for the headphone tests and the regression curve was calculated to be 2.0% of the scores with the coefficient of determination $R^2 = 0.98$.

The results obtained with each method were subjected to an analysis of the variance (ANOVA) for each of the studied scenarios. The null hypothesis, H_0, established that both real-life and headphones samples were similar for each STI scenario. At a 5% level of significance (95% confidence interval), five of the eight STI scenarios presented strong evidence against H_0, (0.36, 0.43, 0.50, 0.54, and 0.60STI, with $p \leq 0.01$), two presented moderate evidence against H_0 (0.40, and 0.56STI, with $0.05 \geq p > 0.01$), and one scenario presented weak evidence against H_0 (0.48STI, with $0.1 \geq p > 0.05$). Table 1 provides summary results for the 10 STI scenarios studied with the real life method and the 8 scenarios studied with the headphone method.

It can be seen from Figure 5 and Table 1 that the headphone method produced lower averaged intelligibility results for all scenarios and higher standard deviations for seven of the eight scenarios.

STI scenario	Real-life		Headphones		Difference in Means	ANOVA	
	Mean	Stdev	Mean	Stdev		p	Evidence against H_0
0.36	54.6	5.4	45.4	8.6	9.2	0.01000	Strong
0.40	64.0	4.2	59.2	3.9	4.8	0.02193	Moderate
0.43	71.6	2.3	63.8	4.8	7.8	0.00034	Strong
0.45	71.0	4.7	NA	NA	-	-	-
0.48	79.0	2.0	75.4	5.9	3.6	0.09939	Weak
0.50	82.8	2.0	73.4	5.6	9.4	0.00034	Strong
0.54	87.0	3.1	80.2	4.0	6.8	0.00028	Strong
0.56	89.2	2.0	86.0	3.6	3.2	0.03516	Moderate
0.60	94.8	2.2	86.8	3.5	8.0	0.00002	Strong
0.70	96.8	1.0	NA	NA	-	-	-

Table 1. Summary results for the real life and headphone listening tests.

3.3 Discussion.

The headphone listening results can be summarised as follows:

- The intelligibility PB-word tests using with headphones produced lower averaged results than those obtained with real-life listening for the eight scenarios.
- The differences between the real-life and the headphones scores were statistically significant for seven of the eight STI scenarios investigated.
- The headphone results showed higher standard deviations than the real-life results for most of the STI scenarios investigated.
- The averaged headphone results showed a higher calculated standard deviation with the best-fit regression polynomial (2.0% of the scores with headphones compared to 1.4% of the scores for real-life listening).

One possible factor contributing to the lower intelligibility scores with headphones is that the headphones may have deprived the listeners of important spatial cues of information which are vital for the suppression of the reverberation.

4. CONCLUSIONS

The following conclusions are made:

- In this study, one hundred subjects were used real-life listening test to investigate the relationship between speech intelligibility using PB word score and measured STI in a noise-free reverberation chamber. One hundred subjects listened to five English word lists in carrier sentences at ten different locations in the chamber, corresponding to ten different STI scenarios.

- Binaural recordings of the PB words were also made at eight locations in the chamber, which were then equalised and presented to eighty listeners over headphones for word score tests.

- The experiments found that the real-life intelligibility test produced higher averaged PB score results than tests in which the speech was presented with headphones.

  The differences between the real-life and the headphone scores were statistically significant for most of the ten STI scenarios investigated. The headphone listening also presented higher standard deviation than the real-life listening for most of the STI scenarios and higher calculated standard deviation of the averaged values from its regression polynomial.

  One possible explanation for the lower intelligibility scores with headphones is that the headphones may have deprived the listeners of important spatial cues of information which are vital for the suppression of the reverberation.

- Despite the binaural nature of the real-life and headphone intelligibility methods used in this study, the measured word scores with both methods were lower than the scores found by Anderson and Kalb for a given STI value. Noting that the Anderson and Kalb relationship between English PB word score and STI is referred by the current STI standard, our results suggest that under reverberant conditions, the subjective speech intelligibility for a specific STI value is more degraded than currently thought.

  One possible factor for the difference between our results and those of Anderson and Kalb is that our listeners were untrained to listen in reverberant environments, where those of Anderson and Kalb had been trained. The use of untrained better represents the experience of the general public in reverberant situations such as transport terminals, churches and sporting facilities.

6. REFERENCES

1. IEC. Sound System Equipment Part 16: Objective rating of speech intelligibility by Speech Transmission Index. 3rd Edition, 2011. International Standard No. 60268-16.
2. Anderson, B.W., and Kalb, J.T. “English verification of the STI method for estimating speech intelligibility of a communications channel,” J. Acoust. Soc. Am. 81, 1987,pp 1982-1985.
3. Leembruggen, G.A, Stacy A. Should the Matrix be Reloaded? Proc IOA. 2003.
4. Mapp, P. Some Effects of Equalisation on Sound System Intelligibility and Measurement. Preprint AES 115th Convention . 2003.
5. Leembruggen, G. Is SII better than STI at recognising the effects of poor tonal balance on intelligibility? Proc IOA. 2006, Vol. 28, Part 6
6. ANSI S3.2-2009, “Method for Measuring the Intelligibility of Speech over Communication Systems”, American National Standards Institute, New York. (2009).
7. Wijngaarden, S, Drullman, R, “Binaural intelligibility prediction based on the speech transmission index” J. Acoust. Soc. Amer. 123, 2008, p.4514-4523.
8. Steeneken, H.J.M. and Houtgast, T., “Validation of the STIr method with the revised model,” Speech Communication 38, 2002, p.413-425.
9. H. MØller, "Fundamentals of Binaural Technology," Appl. Acoust., vol. 36, 1982, pp. 171 – 218.
10. H. MØller, D. HammershØi, C. B. Jensen, and M.F. SØrensen, "Transfer Characteristics of Head-phones Measured on Human Ears," J. Audio Eng. Soc.,vol. 43, (1995 Apr.), pp. 203-217.
11. Kulkarni,A.,and Colburn,H.S.‘‘Variability in the characterization of the headphone transfer-function,’’J.Acoust.Soc.Am.107, 2000, pp. 1071–1074.
12. H. MØller, C. B. Jensen, D. HammershØi, M.F. SØrensen: "Using a typical human subject for binaural recording." Proc. 100th Audio Eng. Soc. Conv. Preprint 4157 (1996) 1-10.