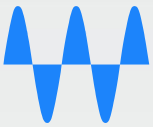

IMPROVING VOICE INTELLIGIBILITY USING WAVES Nx

Amir Ben-Kiki, Matan Ben-Asher, Itai Neoran

Waves Audio Research – Technical Paper

Abstract: Waves Nx is a real-time, head-tracking-based, binaural reproduction system, able to position sound sources in a virtual acoustic space surrounding the listener. This research investigates speech intelligibility in the presence of noise under headphone listening conditions used in typical telecommunications setups. It is shown through subjective experiments that, when using the Waves Nx processing on speech, intelligibility of speech is improved. This paper reviews academic researches by Ben Gurion University (BGU) and by others, concluding on possible applications of Waves Nx in Voice-over-IP applications.

1. INTRODUCTION

Intelligibility of speech in telecommunications in presence of noise and interfering talkers, has been subject to numerous researches. It is well known, for several decades, that in real-world situations, humans possess the ability to focus attention on speech and sounds and separate them from noise and other speech sources, even in noisy environments. This phenomenon is referred to as the "Cocktail Party Effect". It is also known that that the ability to process directional information is a key element in enabling the cocktail party effect. This processing is carried out by cognitive and perceptual interpretation of time delays and level differences between the two ears (binaural information).

Recently, it has become a common use case to listen to speech over stereo headphones, in voice over IP and in other telecommunications applications. This allows a growing field of investigations, seeking to enhance speech intelligibility in a binaural manner.

Waves Nx is a real-time binaural engine, using means for head-tracking (using either Camera Based Face Tracking or IMU sensors) to synchronize the binaural processing with the head position and orientation of the listener. Waves Nx can generate artificial directional information to monophonic sources, and can also reproduce, in a binaural manner, directional information already embedded in stereo or multi-channel input audio streams. By using low latency high-rate head-tracking, Waves Nx allows the listener to explore the virtual world in a precise manner, using conscious and unconscious head movements.

This paper examines the advantages of Waves Nx when used enhance intelligibility of speech over stereo headphones.

Throughout this paper SRT values are provided as negative values, as per industry conventions:

"[SRT is] the minimum intensity in decibels at which a patient can understand 50% of spoken words; used in tests of speech audiometry. Also called speech recognition threshold." – [Mosby's Medical Dictionary]

2. INTELLIGIBILITY USING WAVES Nx IN SINGLE TALKER SCENARIOS

In a recent paper from BGU, referenced in [BGU 2016], the effect of Waves Nx is investigated and measured in a HIST subjective test under multiple test scenarios. The study focuses on measuring the intelligibility of speech in scenarios where either the voice, the interference, or both, are rendered binaurally and spatialized using Waves Nx, listeners were encouraged to take advantage of Waves Nx head tracking capabilities by moving their heads relative to the perceived location of the auditory signals. The study reports statistically significant results, showing SRT (Speech Reception Threshold) score improvements from -1db to -8db depending on the experimental configuration, most significant (-8dB) results are attained when the interference is diotic (monophonic) and the voice is dichotic (binaural), however, a statistically significant -1dB result is measured even when both signals were dichotic and processed through Nx. Figure 1 illustrates the various test configurations tested in BGU experiment.

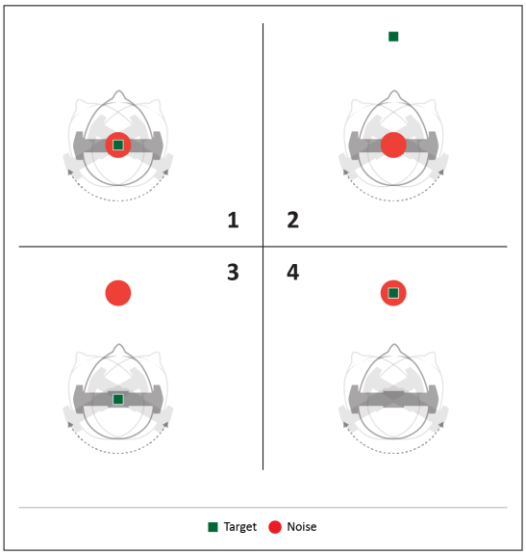

FIGURE 1- BGU EXPERIMENTAL TEST CONFIGURATIONS, SIGNALS APPEARING WITHIN THE LISTENERS HEAD WERE RENDERED DIOTICALLY WITHOUT NX, SIGNALS RENDERED OUTSIDE THE LISTENERS HEAD ARE RENDERED DICHOTICALLY AND EXTERNALIZED USING NX.

Figure 2 shows that, in all three experimental configurations (configuration 1 is a control group where both speech and interference are diotic) application of Waves Nx yields improvement to intelligibility, expressed as the decrease in the SRT. Configuration 4 shows the SRT reduction when both voice and interference are dichotic and reproduced with no directional separation, this aspect of Binaural Reproduction with no directional separation is unique to the BGU study and has not been, to the authors' knowledge at the time of writing, tested nor reproduced using binaural sound reproduction engines other than Waves Nx.

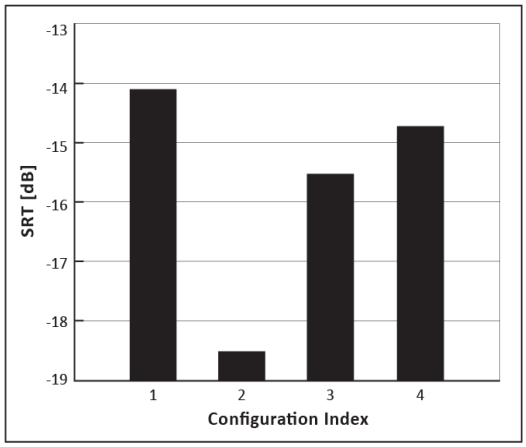

FIGURE 2- SRT CALCULATED FOR ALL TEST SUBJECTS. 95% CONFIDENCE INTERVAL ON VERTICAL BARS.

The superior results achieved with Waves Nx compared to prior literature experiments, can possibly be attributed to the precise tracking of the listener's head orientation, which is incorporated into Waves Nx. Head tracking allows subjects to naturally and intuitively focus on spatialized sound sources, by adjusting the orientation of their heads to face their preferred and optimal head direction towards the virtual sound source, and by removing directional ambiguities using small unconscious head movements.

Prior research [Begault and Erbe 1994] shows that spatial separation between a source and environmental interference dramatically increases voice intelligibility, by up to 6 dB, when source and interference are separated by 90 degrees on the horizontal plane. The latter results, relating to a large spatial separation between the desired speech and the interference, add to the BGU findings, which relate to hardly any directional separation, together covering the two important use-cases in Voice Over IP calls.

3. INTELLIGIBILITY IN MULTIPLE SPEAKER SCENARIOS AND CONFERENCE CALLS

The advantages of binaural reproduction of multiple voice signals for conference calls have been studied and documented as early as 1953 [Cherry 1953], and have been reproduced in both natural as well as virtual environments in multiple studies [Cherry 1953, Yost 1997]. All studies demonstrate that spatial separation of source signals achieved through binaural reproduction, the application of Head Related Transfer Functions (HRTF) and acoustic emulation increase intelligibility during conferences, this phenomenon has been dubbed the "Cocktail Party Effect".

The improvement in speech perception with HRTF results in enhanced intelligibility in presence of noise or several talkers, depending on the angular and spatial separation between the talkers. [Bronkhorst and Plomp 1992] show that, in conditions of 50% intelligibility application of binaural sound reproduction improves SRT scores by -1.5 dB to -8 dB, with results varying in dependence on the noise, the number of talkers and their position. Similar findings are reported by [Ricard and Meirs 1994], with an increase of +5dB in intelligibility when the interference is white noise and positioned binaurally in a frontal position.

Additionally, research shows that further SRT improvement of up to -3.4 dB is achieved through the introduction of three-dimensional separation between sources [Drullmana and Bronkhorst 1999].

Waves Nx allows positioning audio sources in arbitrary positions in the virtual space around the listener which allows taking full advantage of the cocktail party effect, both in planar two

dimensional configurations as well as in spherical three dimensional configurations. Figure 3 is an illustration of a proper Cocktail Party configuration using Waves Nx compared to a monophonic conference call setup: Configuration 1 shows the perceived sound stage in conference call without the application of spatialization, externalization and head tracking; the voice sources perceptually originate from inside the user's head and the Cocktail Party Effect cannot be taken advantage of. Configuration 2 shows the perceived sound stage when the voices are spatialized and externalized using Waves Nx, allowing the listener to take full advantage of the cocktail party effect and to focus on individual sources using Waves Nx's head tracking capabilities.

4. CONCLUSIONS

It has been shown, through several researches and academic reports, that spatial separation of a single source from noise interference yields up to -12 dB decrease in SRT, when multiple sources are present SRT results are improved by up to -8 dB, applicable to Waves Nx. In addition, a recent BGU study shows results specific to Waves Nx that demonstrate -1dB to -8dB SRT improvement even when no explicit directional separation is introduced.

Waves Nx allows taking full advantage of the Cocktail Party Effect using modern software tools that allow reproduction using regular stereophonic headphones. It also allows 3D separation between speech sources, which further increases intelligibility in conferencing scenarios.

Waves Nx uses precise head tracking, and its advantages for speech processing stem both from the static spatial separation, and from spatial information embedded in expected audio changes corresponding to natural head movements, allowing the listener to intuitively focus the attention towards specific spatialized sources, in similar mechanisms as in real world conversations.

5. REFERENCES

[BGU 2016] - “Intelligibility of speech in noise under diotic and dichotic binaural listening”, Noam R. Shabtai and B. Rafaely, Ben Gurion University of the Negev, Nov 20, 2016.

[Mosby's Medical Dictionary, 9th Edition] – Mosby, Elsevier Health Sciences, 29 Apr 2016 p. 1662

[Begault and Erbe 1994] – Begault, D. R., and Erbe, T. 1994. ‘‘Multichannel spatial auditory display for speech communication,’’ J. Audio Eng. Soc. 42, 819–826.

[Cherry 1953] – Cherry, E. C. 1953, ‘‘Some experiments on the recognition of speech, with one and with two ears,’’ J. Acoust. Soc. Am. 25, 975–979.

[Yost 1997] - Yost, William A. "The cocktail party problem: Forty years later." Binaural and spatial hearing in real and virtual environments (1997): 329-347.

[Bronkhorst 2000] Bronkhorst, Adelbert W. "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions." Acta Acustica united with Acustica 86.1 (2000): 117-128.

[Drullmana and Bronkhorst 1999] - “Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation” - Rob Drullmana and Adelbert W. Bronkhorst, Acoustical Society of America, TNO Human Factors Research Institute 1999.

[Bronkhorst 2000] Bronkhorst, Adelbert W. "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions." Acta Acustica united with Acustica 86.1 (2000): 117-128.