VOICE BALANCE — A SPIRIT LEVEL BASED ON VOCAL SOUNDS

Piet Kuchenbecker

University of Hamburg
Institute of Systematic Musicology
Hamburg, Germany
piet.kuchenbecker@studium.uni-hamburg.de

ABSTRACT

Sonification is informative. But a weakness of sonification utilizing artificial sounds is that people are not familiar with it and get annoyed quickly. Also, most of the information mediated by sound can hardly be understood intuitively. In Voice Balance, you listen to the human voice to level your smartphone. The sound is familiar, informative, and uses sounds found in our language to communicate.

1. LINK TO APK FILE

The APK file of Voice Balance can be found under `https://github.com/Tiltification/sonic-tilt/tree/voicebalance/build/app/outputs/flutter-apk`.

2. INTRODUCTION

The psychoacoustic sonification in Tiltification [1] is a highly unnatural sound. Shepard tones [2] do not appear in nature. They even create auditory illusions. It is no wonder that some users have troubles accepting this as a source of information [3]. The human voice is a natural and familiar alternative.

The sounds used in Tiltification are not intuitive to humans. A user of the app has to read a manual to understand the information the app is giving, and it can be hard to remember. An ascending or descending Shepard tone is not intuitively associated with left or right. This makes a classical bubble level more practical to most users. The spoken words "left" and "right" are sounds our brain associates with the relative directions left and right, but spoken words are not practical for this use. They are momentary and can only transport information for a short period of time. Information can be transferred over a longer time span using short looped vocal chops of the spoken words. In the context of a level, a long "l-" or "r-" sound can still be associated with left and right.

Using ridiculously fast voice samples to convey information is known as *spearcons* [4]. A strength of spearcons over speech at normal speed is that spearcons are more time-efficient. That is the reason why blind and visually impaired people tend to utilize screen readers at high speed [5]. Another benefit of spearcons is that it is helpful, but not necessary, that you understand the spoken words. Spearcons have a unique timbre that can be learned in various contexts. Just as for auditory icons, the downside of spearcons is that they are suitable to display discrete events and nominal data rather than continuous data [6]. This makes them unsuitable for a spirit level.

Alternatively, a number of voice-like sonifications have been invented [7, 8, 9]. Voice Balance follows this strategy to allow users to level their phone by interpreting the meaning of the human voice. Guidance by human voice is intuitive and already learned as a child.

3. VOICE BALANCE

A short snippet of recorded human voice is the core of Voice Balance. Such an approach is referred to as *sample-based display* [10]. The two-dimensional space is divided into four phonetic directions [l], like "l"eft, [r], like "r"right, [u], like t"u"be ("u"nten in German, which means *down*) and [o], like b"o"at ("o"ben in German, which means *up*). How much the smartphone is tilted in that direction is expressed by the gain of the particular phoneme. This principle is illustrated in Fig. 1. The gain function is

$$g(x) = \sqrt{|x|} \tag{1}$$

for the [l] or [r]sample, and

$$g(y) = \sqrt{|y|} \tag{2}$$

for the [u] or [o] sample. The function is illustrated in Fir.g 2. This nonlinear function implies that the gain changes more dramatically at small tilt angles, raising the precision near the optimal tilt level. This is meaningful, because users need to hear whether they are 1° or 3° off, but not whether they are 40° or 42° off.

The left/right dimension uses a consonant and one determined pitch. The up/down dimension uses a vowel sound and two different pitches for up and down, to separate them, despite the acoustic similarities of the vowels [u] and [o]. The [o]'s pitch is a fifth below and the [u]'s pitch, and a minor third below the left-right-axis, which leads to harmonic sounds, if both axes aren't centered. With these differences, users can always listen to both dimensions at the same time and tell them apart.

There are two regions near the target $(0^\circ, 0^\circ)$ point. Both trigger background noise. The outer region indicates that the phone is almost leveled. But the noise is so quiet that you can still hear the voice. This allows for final corrections of the tilt angles. The inner region triggers additional noise, while the human voice keeps getting quieter, the more you approach the target point. This way

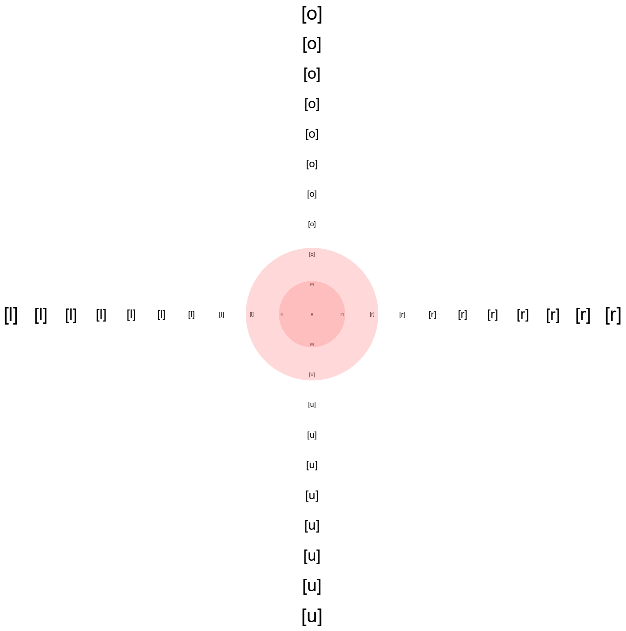

Figure 1: *When tilting in either direction, the corresponding consonant ([l] or [r]) and/or the corresponding vowel ([o] or [u]) are being played. The larger the tilt angle, the higher the gain. Near the target, pink noise is triggered as a confirmation sound. Even nearer, a second pink noise generator starts playing. Together, they mask the voice sample, so users do not hear a wild fluctuation near perfect level.*

the noise masks the phonemes so that you do not hear sound fluctuations of marginal tremor near the perfect level.

This approach of a harmonic sound prevailing against a background noise is related to the signal-to-noise ratio approach [11] that is sometimes used as a sonification strategy. Here, the data magnitude is mapped to the sound pressure level of a harmonic sound in presence of a background noise.

Complex amplitude and frequency modulations are implemented to simulate jitter and shimmer [11], which makes the appearance of the voice more dynamic and natural.

Note that the timbre of [l] and [r] can be distinguished well. Even if you do not associate these consonants with directions in your native language, you can easily learn their meaning inside the app.

The same is true for [u] and [o]. These vowels can be distinguished well. However, to increase the difference between them, they received different pitches in Voice Balance.

Telling the two dimensions apart is easy, too. You can decide to concentrate either on the consonant or on the vowel, as they have distinguishable timbres and pitches.

The application consists of four sample players whose amplitudes are being altered by the output values of the devices tilt sensors. The x-axis controls the “l” and “r” sounds and the y-axis the “u” and “o” sounds. The two parameters outputted by the tilt sensors have a value of zero once the device is centered. When centered, every sample’s amplitude is multiplied by zero which leads to all the four samples being muted. Only one sample per axis can be outputted at a time. This is made possible by outputting to equal values, multiplying one by -1 and taking the square root of both of them. The square root of a negative number is mathematically imaginary, which causes the PD-code to output a zero. The square root of the positive value can be calculated and outputted.

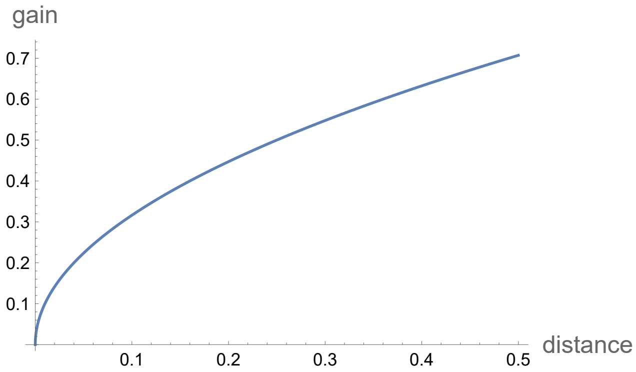

Figure 2: *The gain function is a square-root-function. It produces larger gain changes near the target.*

Once the output values of the “x” and the “y” axes are below a certain threshold and close enough to zero a noise generator activates. When the device is centered, the samples will be muted, but the pink noise will be played to indicate the phone being centered.

Each sample player has its own shimmer and jitter effects with unique values for variation.

4. REFERENCES

[1] M. Asendorf, M. Kienzle, R. Ringe, F. Ahmadi, D. Bhowmik, J. Chen, K. Huynh, S. Kleinert, J. Kruesilp, Y. Lee, X. Wang, W. Luo, N. Jadid, A. Awadin, V. Raval, E. Schade, H. Jaman, K. Sharma, C. Weber, H. Winkler, and T. Ziemer, “Tiltification — an accessible app to popularize sonification,” in *Proc. 26th International Conference on Auditory Display (ICAD2021)*, Virtual Conference, June 2021, pp. 184–191. [Online]. Available: https://doi.org/10.21785/icad2021.025

[2] R. N. Shepard, “Circularity in judgments of relative pitch,” *The Journal of the Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, 1964.

[3] T. Ziemer and N. M. Jadid, “Recommendations to develop, distribute and market sonification apps,” in *The 27th International Conference on Auditory Display (ICAD 2022)*, Virtual Conference, June 2022. [Online]. Available: https://doi.org/10.21785/icad2022.003

[4] B. N. Walker, J. Lindsay, A. Nance, Y. Nakano, D. K. Palladino, T. Dingler, and M. Jeon, “Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus,” *Human Factors*, vol. 55, no. 1, pp. 157–182, 2013, pMID: 23516800. [Online]. Available: http://dx.doi.org/10.1177/0018720812450587

[5] B. Biggs, J. M. Coughlan, and P. Coppin, “Design and evaluation of an audio game-inspired auditory map interface,” in *25th International Conference on Auditory Display (ICAD 2019)*, Newcastle upon Tyne, UK, June 2019. [Online]. Available: http://hdl.handle.net/1853/61525

[6] T. Ziemer, N. Nuchprayoon, and H. Schultheis, "Psychoacoustic sonification as user interface for human-machine interaction," *International Journal of Informatics Society*, vol. 12, no. 1, pp. 3–16, 2020. [Online]. Available: http://www.infsoc.org/journal/vol12/12-1

[7] D. Rocchesso, S. Andolina, G. Ilardo, S. D. Palumbo, Y. Galluzzo, and M. Randazzo, "A perceptual sound space for auditory displays based on sung-vowel synthesis," *Scientific Reports*, vol. 12, p. 19370, 2022.

[8] T. Hermann, G. Baier, U. Stephani, and H. Ritter, "Kernel Regression Mapping for Vocal EEG Sonification," in *Proceedings of the 14th International Conference on Auditory Display (ICAD2008)*, Paris, France, June 2008. [Online]. Available: http://hdl.handle.net/1853/49939

[9] F. Grond and T. Hermann, "Singing function," *J Multimodal User Interfaces*, vol. 5, pp. 87–95, 2012. [Online]. Available: https://doi.org/10.1007/s12193-011-0068-2

[10] G. Dubus and R. Bresin, "A systematic review of mapping strategies for the sonification of physical quantities," *PLOS ONE*, vol. 9, no. 4, p. e96018, 12 2013.

[11] D. Arfib, J. Couturier, K. L., and V. Verfaille, "Strategies of mapping between gesture data and synthesis model parameters using perceptual spaces," *Journal of Organised Sound*, vol. 7, no. 2, pp. 127–144, 2002.