

Metrics for Electronic data Representation into the Audio Space

Dominik Ďurikovič¹ and Roman Ďurikovič^{1,2}

¹ University of Saint Cyril and Metod, Trnava, Slovakia,
d.dominik@zoznam.sk

² Faculty of Mathematics, Physics and Informatics,
Comenius University, Slovak republic
roman.durikovic@fmph.uniba.sk

Abstract. We discuss the representations of the electronic data, especially content of the web pages, into the audio space. We define two metrics usable for comparing two distinct audio representations. First metric is connected with data changes perception by reading users. Changes in the web page that are not important for the reader comparing to the previous version are called small changes on the web page. Changes with higher importance value for reader are called huge changes. Ability to catch relevant changes in source data document by audio space representation is important measurable attribute for users reading these data. This attribute is called web change perception for web pages. First, we describe evaluation of the web change perception and next we are measuring sound perception of these representations. The applications of the metric are in tuning interfaces for representing web pages into the audio space.

1 Introduction

Most electronic data representations are designed for people who see well. We consider devices such as LED diodes, displays on PDA handhelds or LCD displays for representation of electronic content. Group of visually disabled users, or another group of users with busy sight (car drivers) can not operate these devices properly to retrieve the represented information.

Web page representations of the largest bank of information, the Internet, become an important task for people. Visually impaired users have lot of problems with electronic data representation in a proper way. As the visually impaired people say, most of their Web site problems are caused by images, forms, tables, Javascript and maps because many Web pages are not created in accordance with standard guidelines (mainly formulated by World Wide Web Consortium - W3C).

Within this paper we build basis for quality measures of the new mappings to represent importance of content and navigation information to visually disabled users using auditory channels. The main focus of our research is on a novel control and display mechanism using 3D auditory channels on PDA. Our results

can also help to restructure or project the content of Web pages for small screen displays. Such mappings can allow a user to follow the Web content projected to 3D auditory channels while performing other actions. For example, car drivers could also use the results of this research by browsing the Web with a few bottoms on their driving wheel.

The basic core of our electronic content representations will be a 3D audio space. Elements of the 3D audio representation space are tones, synthetic sounds, environmental sound and spoken texts.

2 Related Work

There are several projects that try to solve the problem of web page content representation for visually impaired people, to give the user a general overview of the web page content and to make the work with information easier. In the project Air-Client, see [1], individual parts of pages are represented with the assistance of spatial sound by creating of hearcons in auditory interaction realm - Air. Project [2] is aimed to provide visually impaired internet user with spatial and navigation information with the help of speech, sounds and haptic device. Another method [3] has been developed to identify small, common interaction design problems, and design and evaluate several solutions for each.

3 WEB representation in 3D-ARWEB

We call *3D audio representation of the internet Web page (3D-ARWEB)* mapping projecting source of the Web page into the three dimensional audio space. Let's start talking about the way we may represent data using audio space (3D-ARWEB). We decided to place sound sources to horizontal plane only and most of sounds will be in the front half-plane of the head. This is because there are many front-back and top-down confusions reported in previous researches by users. So sound sources will be placed at reduced semi-circle, called Stage-Arc shown on Figure 1. The reduction is caused by lower accuracy of human's ear to azimuth change in both extreme positions.

We use four different synthesized male voices (speakers). First reads headings, second reads contents, third is for links and the last one belongs to images. Reading images means to read their alternative description. It is up to web developer to provide such information. All voices are located at the Stage-Arc right in front of listener. They don't move, because we think that it is not pleasant when a text is read from one side for a long time. The reason why to use more voices is that human can easily recognize which voice belongs to which speaker. It is sufficient just to read a text and user immediately knows if it is a heading or a link. Actually we use Festival open source project to synthesize speech. We are restricted to use four male voices as they are provided with Festival. To create more voices is out of scope of our project.

Along with synthesized speech we use non-speech sounds called earcons. An earcon is any sound that does not contain human voice neither synthesized nor

prerecorded. It is a sound symbol with exact meaning. Four earcons are used with the same meaning as voices. So we have one earcon for headings, one for contents, another for images and the last is for links. When a text is being read then earcon is sounded from the relative position along the Stage-Arc. It provides information about position in document. Its position on the Stage-Arc is relative to position in document.

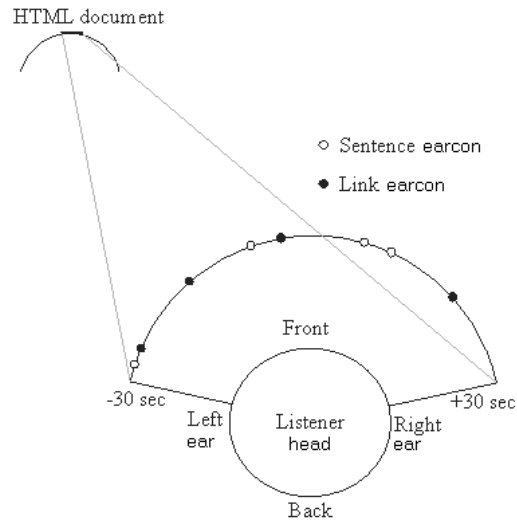


Fig. 1. 3D-ARWEB sound survey of a specific region of the HTML document.

One of basic problems in hypermedia systems for visually impaired is the way how links are represented and followed. Links are of two types: intra-document and inter-document. If these two links are not represented in different ways, user can get lost in hyperspace. The way we solved this problem follows. Intra-document link uses three sounds: take off, flying and landing sound. When user activates this type of link the take off sound is played from the relative position on the Stage-Arc. Then flying sound moves towards target and at last landing sound is played.

For inter-document link we use only two sounds: take off and landing. These are different from sounds used in intra-document link. Take off sound fades out and its distance from listener grows. Landing sound comes nearer and becomes louder.

Another problem is activating a link. We mark a link as active when it is being read. To mark link as active is only internal action of the browser. No special sound is played. It remains active until another link is encountered in the document. User can follow active link at any time by pressing space bar on keyboard.

When sighted users browse, they firstly create quick overview of whole document by looking at it. This is what blind users cannot do. We provide this functionality by creating a sound survey of the document surrounding the user's current position in the document. Blind users may focus in detail on specific part of the document after sound survey.

Representing web page in 3D-ARWEB is possible in four layers. In the first layer whole document is read and objects are represented as earcons. The second layer reads and represents the links. The third represents the headings and the fourth reads only textual content without other structural elements.

Two different representations of the same electronic content may vary in the perception by users. Due to this a lots of tests on our representations are needed to tune them and find out the best one. The idea is to construct perception space of our 3D-ARWEBs and specific set of the metric functions to evaluate distances between our representations. The structure of the measures for 3D-ARWEB are:

- sound quality
- sound space perception by users (hearable frequencies, low sound masking, loudness, localization, ...)
- speed of the data mining in current representation by user
- stability of the representation

Following sections are focused on specification of metric elements for quality measurement of the 3D-ARWEB space based on web change perception. Next section construct sound space perception for 3D-ARWEB.

4 Web change perception

We say that proposed 3D-ARWEB *is stable* if small visible change of the source Web page implies small hearable (data, or structure) changes in the 3D audio representation. Changes of the Web changes are calculated by percent ratio between quantity of changes concerning to the whole Web page.

We would like to measure stability of the 3D-ARWEB via perception of the web page changes between classical Web representation on the LCD display and our 3D-ARWEB in our research.

One issue when studying the human perception of changes is that there are many possibilities to change the Web page. Therefore, we need to classify the nature of web changes. A possible structure for changes is to classify them as

- content
- presentation
- structural
- and behavioral changes.

Content changes refer to modifications of the page contents from the reader's point of view. For example, a Web page that contains information about a toy company's products might be continuously updated by the prices of toys. Later,

before Christmas there will be detail information about one great Christmas action on one specific toy.

Presentation changes are changes related to the visual representation of material that do not change the material itself. Examples of presentation changes are different design layouts, fonts, backgrounds or styles of bullets.

Structural changes refer to the underlying connection of the Web page to other Web pages. These connections are expressed by web links. Structural changes may not result in any visually perceptible change to a Web page except when the mouse moves over the links.

Behavioral changes refer to the modifications in web page code not resulting in any visually perceptible change.

In the proposed 3D-ARWEB we signify mostly web visually perceptible changes content changes and presentation changes.

Some work on the web change perception on few sample tests is described in the work [4]. The web pages were partially changed in a few percent and users were asked to locate specific class of changes during limited time. The results of these tests show dependence of perception on time measured for several sample web changes. We may run similar tests for web change perception in 3D-ARWEB space with visually impaired users and compare the results to the web change perception on LCD displays with healthy users.

Differences between Web changes perception on LCD display and 3D-ARWEB defines metric expressing similarities in perception between classical visual Web representation and 3D-ARWEB. This metric can be calculated after several tests for current 3D-ARWEB. Metric shows a small differences in perception space when we have small changes between the Web page source and the 3D-ARWEB space.

Due to this fact we can say that the stability of the 3D-ARWEB is measured by the web change perception metric.

5 Basic sound cognition

The 3D-ARWEB is an audio interface and due to this the sound perception is another element affecting metric of the 3D-ARWEB space.

Elements of the sound space has a few physical characteristics like loudness, sound masking, pitch and sound position.

Loudness is the intensity attribute of an auditory sensation, in terms of which sounds may be ordered on a scale extending from quite to loud [5]. The sound pressure level in dB of a pure tone of frequency 1kHz which is judged by the listener to be equivalent in loudness is the loudness level.

Masking is the amount by which the threshold of audibility for one sound is raised by presence of another (masking) sound.

Pitch is that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale.

Localization of the sound position refers to judgments direction and distance of a sound source.

Every sound is represented with the spectrum. The spectrum of a sound wave is the distribution in frequency of magnitudes of components of the wave. It can be represented by plotting power, intensity, amplitude, or level as function of frequency [5].

6 Sound perceptual measure

We say that 3D-ARWEB is *better perceptable* than another one 3D-ARWEB while the sound masking is lower and differences between sounds used in the 3D-ARWEB are higher. Sound recognition is not well studied yet.

The sound perception and cognition researcher clarified catch-all term *Timbre* as in [6]. This term represents all aspects of a sound, independent of pitch and loudness. Most quantitative approaches to timbre perception describe the distance between two sounds. It is often described by a combination of subjective perceptual dimensions. There is no principled way to synthesize timbres which will lead to a prescribed perception.

The work [6] describes timbre like isomorphic representation with human perception. They created a perceptual space that describes the connection between physical attributes of a timbre and human perception. This model is called as a timbre space. Such a model is vital for timbre-based auditory display.

Let's specify sound representation for timbre space creation. Each spectrum is unique representation of the specific sound. Spectrum forms complete sound representation and its arbitrary complexity makes a direct mapping to human perception unknown. The complexity of the sound spectrum is the reason why in [6] is Mel-frequency Cepstral Coefficients (MFCC) taken as good enough representation of a sound. In this case the spectral shape is represented by 13 ordered coefficients C_i . Average power in the spectrum is represented by first coefficient and the broad shape of the spectrum represents second coefficient. Higher order coefficients represent inner details of the spectral shape.

6.1 MFCC calculation

In the MFCC, the frequency amplitude bands are positioned logarithmically which approximates the human audio perception more closely than the linearly-spaced frequency bands obtained from the fast fourier transform or discrete cosine transform (DCT) only. As mentioned in [7] MFCC is derived using the following steps (Figure 2)

- Take the Fourier transform of a sound spectrum $S(f)$. The frequency is warped according to the critical bands of human hearing.
- Filterbank $H_i(f)$ has the passband of 133.3Hz for the first 13 channels between 0Hz and 1kHz and wider. Band width is expressed like

$$BandWidth(H_i) = \begin{cases} 133.3 & (i \leq 13) \\ 1000 \cdot 1.072^{i-13} & (i > 13). \end{cases}$$

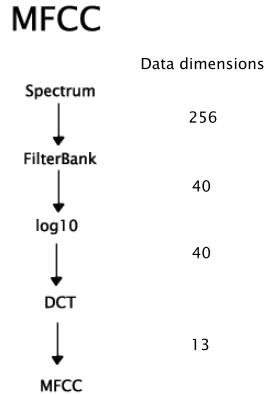


Fig. 2. Sound representation using MFCC.

Total energy in each channel F_i is integrated to express filterbank output.

$$F_i = \int |H_i(f) \cdot S(f) df|$$

where i is the channel number in the filterbank, and $H_i(f)$ is the filter response of the i th channel.

- MFCC coefficient C_i are computed using discrete cosine transformation ([6])

$$L_i = \log_{10}(F_i)$$

$$C_i = DCT(L_i)$$

There can be variations on this process e.g. differences in the Mel scale conversion. Popular approaches of this sound representation are based on speech perception, speech recognition, and the perception of musical sounds.

MFCC represent a timbre as a low dimensional vector. Any point in this multidimensional space is a sound, which for visual purposes we can display as a spectrogram. Sound representation using MFCC is perceptually orthogonal. This means that changes in one parameter (coefficient C_i) do not affect perception of the other axis (coefficient C_j where $i \neq j$).

Timbre is a multidimensional quantity and an important metric in this work is that the representation's axis be perceptually orthogonal. This means that changes in one parameter do not affect perception of the other axis.

6.2 Inverse transform of the MFCC

The spectral shape reconstruction from MFCC starts with DCT inversion on coefficients C_i and amplitude scaling

$$\tilde{L}_i = IDCT(C_i)$$

$$\tilde{F}_i = 10^{\tilde{L}_i} .$$

Assume that \tilde{F}_i represents the value of reconstructed spectrum $\tilde{S}(f)$ at the center frequency of each filter bank

$$\tilde{S}(cf_i) = \tilde{F}_i$$

where cf_i represents center frequency of the i th auditory filter. We have discrete values of the spectrum represented by cf_i at this point. Therefore we will interpolate these values to obtain whole smooth spectrum .

The IDCT of the MFCC is a smooth version of the filterbank output, discarding the fundamental frequency and its harmonics. The reconstructed spectrum has a smooth spectral shape.

6.3 3D-ARWEB sound perception space and metric

As mentioned in the previous text we will represent our sounds used in the 3D-ARWEB as set of the 13 coefficients from MFCC sound representation. These MFCC vectors of coefficients defines space of sounds (MFCC space). Spectrum of the sound in confrontation with MFCC coefficients is unique identifier of the sound. Due to the discretisation (processed filters and transforms) one MFCC vector defines group of very similar sounds. We can expect that one MFCC vector represents group of similar sounds without perceived differences using human hearing.

The perception orthogonality of the MFCC space implies that we can use standard Euclidean metric in this space with 13 dimensions. Distance between two sounds are distance between points calculated using MFCC representation for these sounds.

Next step we need to define space of the audible perceptible 3D-ARWEB representations. All sounds used in 3D-ARWEB are expressed by MFCC 13 dimensional vector. Then we have a finite set of MFCC vectors for every 3D-ARWEB representing perception of the 3D-ARWEB element in 3D-ARWEB sound perception space. Perception of one 3D-ARWEB should be possibly expressed by one function. We express perception of current 3D-ARWEB presentation by function

$$f(\mathbf{x}) = a_0 + a_1 \cdot x_1 + \dots + a_{12} \cdot x_{12}$$

where this function is closest twelve dimensional space to the all sounds from current 3D-ARWEB represented as MFCC (points). To find out this function we need to solve minimum least square error for current 3D-ARWEB

$$\sum_{i=1}^n [y_i - f(\mathbf{x}_i)]^2 ,$$

n is sound count in the 3D-ARWEB, y_i is 13th value of the MFCC vector for i th sound. The vector \mathbf{x}_i consists of first twelve values of the MFCC vector for i th

sound. Suppose that this expression equals zero. If you get first derivatives by a_k (for all k) (unknown const. yet) the equation will be as follows:

$$\sum_{i=1}^n y_i = \sum_{i=1}^n (a_0 + \sum_{j=1}^{12} a_j x_{ij})$$

$$\sum_{i=1}^n x_{ik} y_i = \sum_{i=1}^n x_{ik} (a_0 + \sum_{j=1}^{12} a_j x_{ij}) .$$

Therefore final equations can be expressed as the following matrices

$$X\mathbf{y} = XX^T \mathbf{a}$$

$$\mathbf{a} = (XX^T)^{-1} X\mathbf{y}$$

Thus we finally found coefficients a_i of the function $f(\mathbf{x}) = a_0 + a_1 \cdot x_1 + \dots + a_{12} \cdot x_{12}$ representing sound perception of the current whole 3D-ARWEB. Let's call this function *global audio perception function of the 3D-ARWEB representation*.

We say that two 3D-ARWEB representations are similar in sound perception while their global audio perception functions f_1 and f_2 satisfy condition

$$\int |f_1 - f_2| < \varepsilon$$

where ε is small predefined constant. This integral is also taken as metric on the 3D-ARWEB space.

The *thickness of 3D-ARWEB* is defined as farthest distance between all MFCC sound representations in 3D-ARWEB. The *skimpiness of 3D-ARWEB* is defined as closest distance between all MFCC sound representations in 3D-ARWEB.

We say that 3D-ARWEB representation A is better in sound perception than 3D-ARWEB representation B if and only if skimpiness of A is higher than skimpiness of B .

7 Conclusion

In this paper we have defined so called 3D-ARWEB representation of the web page representation into the audio space. We have defined stability of the 3D-ARWEB representation via the web changes. This stability term implies metric between classical Web page providing and 3D-ARWEB. The new audio perception space of the 3D-ARWEB representations was constructed to measure audio perception between 3D-ARWEB representations.

Finally, we have not begun to take in mind sound and earcons localization in our metrics yet. The web changes perception needs work on tests with visually impaired users to mathematically express 3D-ARWEB change metric .

Our work aims to understand perception of different data representation and to construct metric for comparing and evaluating these 3D-ARWEB representation.

8 Acknowledgements

This research was sponsored by European grant from EU-FP6-MC-040681-APCOCOS and partially supported by VEGA-1-3083-06.

References

1. Donker, H., Klante, P., Gorny, P.: The design of auditory user interfaces for blind users. In: NordiCHI '02: Proceedings of the second Nordic conference on Human-computer interaction, New York, NY, USA, ACM Press (2002) 149–156
2. Murphy, E., McAllister, G., Strain, P., Kuber, R., Yu, W.: Audio for a multimodal assistive interface. In: In proceedings of ICAD'05 Workshop. (2005) 19–24
3. Thornton, C., Kolb, A., Gemperle, F., Hirsch, T.: Audiocentric interface design: A building blocks approach. In: Proceedings of the 2003 International Conference on Auditory Display, July 6-9. (2003) 1–4
4. Francisco-Revilla, L., III, F.M.S., Furuta, R., Karadkar, U., Arora, A.: Perception of content, structure, and presentation changes in web-based hypertext. In: Hypertext. (2001) 205–214
5. Moore, B.C.J.: An Introduction to the Psychology of Hearing. 5th edn. Elsevier Academic Press (2004)
6. Hiroko Terasawa, M.S., Berger, J.: Perceptual distance in timbre space. In: Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display. (2005) 61–68
7. Fang Zheng, G.Z., Song, Z.: Comparison of different implementations of mfcc. J. Computer Science and Technology **16**(6) (2001) 582–589