Tsonos and Kouroupetroglou *EURASIP Journal on Audio, Speech, and Music Processing* (2016) 2016:8
DOI 10.1186/s13636-016-0087-8

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Prosodic mapping of text font based on the dimensional theory of emotions: a case study on style and size

Dimitrios Tsonos and Georgios Kouroupetroglou*

Abstract

Current text-to-speech systems do not support the effective provision of the semantics and the cognitive aspects of the documents' typographic cues (e.g., font type, style, and size). A novel approach is introduced for the acoustic rendition of text font based on the emotional analogy between the visual (text font cues) and the acoustic (speech prosody) modalities. The methodology is based on: a) modeling reader's emotional state response ("Pleasure", "Arousal" and "Dominance") induced by the document's font cues and b) the acoustic mapping of the emotional state using expressive speech synthesis. A case study was conducted for the proposed methodology by calculating the prosodic values on specific font cues (several font styles and font sizes) and by examining listeners' preferences on the acoustic rendition of bold, italics, bold-italics, and various font sizes. The experimental results after the user evaluation indicate that the acoustic rendition of font size variations as well as bold and italics is recognized successfully, but bold-italics are confused with bold, due to the similarities of their prosodic variations.

Keywords: Text-to-speech, Text signals, Typographic cues, Document accessibility, Emotions, Expressive speech synthesis, Document-to-audio, Typographic profile

1 Introduction

Written documents, either printed or electronic, include books, journals, newspapers, newsletters, gazettes, reports, letters, e-mails, and webpages. According to McLuhan, a document is the "medium" in which a "message" (information) is communicated [1]. With the term text document, we refer to the textual content only of a document. A text document contains a number of presentation elements or attributes that arrange the content on the page and apply design glyphs or typographic elements (i.e., visual representation of letters and characters in a specific font and style). For example, the title of a chapter can be recognized as a sentence or phrase placed at the top of the page and in larger font size than the body of the text. Moreover, text color or the bold font style can be used to indicate emphasis in a specific part of a text document. In general, typographic attributes or cues constitute features of text documents, including typeface choice, size, color, and font style. Lorch [2] introduced the term "signal" as the "writing device that emphasizes aspects of a text's content or structure without adding to the content of the text". Text signals "attempt to pre-announce or emphasize a specific part of a document and/or reveal content relationship" [3, 4]. Headings or titles in text documents are considered as signals [5]. Moreover, "input enhancement" is an operation whereby the saliency of linguistic features is augmented through textual enhancement for visual input (i.e., bold) and phonological manipulations for aural input (i.e., oral repetition) [6]. Typographical elements can be conceptualized as semiotic resources for authors, illustrators, publishers, book designers, and readers to draw upon to realize textual or expressive meanings in addition to interpersonal and ideational meanings [7].

Focusing on the visual presentation as well as the organizational aspects of text documents, Tsonos and Kouroupetroglou [8] identified the following:

* Correspondence: koupe@di.uoa.gr
Department of Informatics and Telecommunications, Speech and Accessibility Lab, National and Kapodistrian University of Athens, Panepistimioupolis, Ilissia, GR-15784 Athens, Greece

1. Logical layer: associates content with architectural elements such as headings, titles/subtitles, chapters, paragraphs, tables, lists, footnotes, and appendices.
2. Layout layer: associates content with architectural elements relating to the arrangement on pages and areas within pages, such as margins, columns, and alignment.
3. Typography layer: includes font (type, size, color, background color, etc.) and font style such as bold, italics, and underline.

In contrast to the rich text, the term plain text indicates a text document in a unique font type and size, but without font style.

The abovementioned three layers are complementary and not independent. Typography can be applied to both the logical and the layout layers of a document. For example, a footnote (logical layer) can be a sentence or paragraph in italics or in smaller font size than the body of the text. The vertical space in a text block, called leading (layout layer), can be affected by the font type. Moreover, typography can be applied to the body of the text directly. For example, a word in bold can be used either for the introduction of a new term, to indicate a person's name, or a sentence in bold can be the definition of a term. In this work, we study the typography only; thus, the other two layers (logical and layout) are ignored.

All text signaling devices, either mentioned as typographic attributes/cues, signals or layers, according to Lorch [2]: "a) share the goal of directing the reader's attention during reading, b) facilitate the specific cognitive process occurring during reading, c) ultimate comprehension of text information, d) may influence memory on text and e) direct selective access between and within texts".

The organization of a document can be classified into two main aspects: the logical and the physical. The logical layer of the document defined above corresponds to its logical organization with the same elements (e.g., headings, titles/subtitles, chapters, paragraphs, tables, lists, footnotes, and appendices). At the page level, the physical organization of a document is described by its layout layer in connection with the physical realization of a number of logical layout elements. The organization of a printed or an electronic multipage document as a whole corresponds with the physical implementation of a number of logical layer elements (e.g., chapters, appendices, and indexed references). The organization of a document is domain-specific (e.g., text book, scientific paper, technical report, newspaper, and magazine). The authors use typography and layout in a specific way, e.g., they have to follow strict typographic rules for the documents to be published in a scientific journal. But, in the case of newspapers and books, the page designer (or the page manager), and not the author, has the primary responsibility for applying the typography and the layout layers.

Persons with print disabilities (i.e., individuals who cannot effectively read print because of a visual, physical, perceptual, developmental, cognitive, or learning disability [9, 10]), the elderly, as well the moving user, require printed or electronic documents in alternative formats, such as audio, braille or large print. Text-to-speech (TtS) is a common software technology that converts in real-time any electronic text into speech [11]. It can be combined with other assistive technology applications, such as screen readers, to provide document accessibility through the acoustic modality to those with print disability. Although TtS is considered a mature technology, current TtS systems do not include effective provision of the semantics and the cognitive aspects of the visual (e.g., typographic attributes) and non-visual (e.g., logical layer) text signals [12].

Document-to-audio (DtA) belongs to the next generation of the TtS systems [13], supporting the extraction of the semantics of document metadata [14] and the efficient acoustic representation of text formatting [15–17] by: a) combining alternative text insertion in the document text stream, b) altering the prosody, c) switching between voices, and/or d) inserting non-speech audio (like earcons, auditory icons, or spearcons) in the waveform stream, according to the class of metadata extracted from the document.

Previous approaches for rendering typography to auditory modality can be characterized as direct mapping methodologies. Most of them are based on the relation similarity, i.e., each typographic attribute is directly mapped into a respective acoustic cue. The principle of relational similarity explores two physical quantities with magnitudes that humans perceive by different senses in an analogous way. For example, the font size of a text and the volume of the speech signal when the text is vocalized comprise relational similarity in the case we perceive the change of their magnitudes in a proportional way. In previous studies, the bold typographic attribute is rendered with: verbal description (the phrase "in bold" is said before the salient word with a 15 % decrease of the current pitch) [18], increase (13 %) of the default pitch for each pronounced salient word [18], a two-semitone decrease of pitch voice [19], slower speed for individual words [20], and a ring of a bell before a word with emphasis [21]. The italics typographic attribute is rendered either with a small change in the rhythm of speech [22] or by mixing a sound by 45 % to the right in stereo speakers [19].

W3C introduced in 2012 the speech module [23] for defining the properties of the aural cascade style sheets [24] that enable authors to declaratively control the rendering of documents via speech synthesis or using optional audio

cues. But, both of them are still draft documents, and their publication as a candidate recommendation does not imply endorsement by the W3C. Moreover, although they can be used for direct mapping of typographic cues to corresponding speech properties, they do not define explicitly the required relations for these mappings. For example, they do not provide any information which specific speech properties and how much you have to modify in the case of the "strong emphasis" element which corresponds to the bold font style.

Through a number of psychoacoustic manipulations (pitch, volume, and speed variations of synthetic speech), Argyropoulos et al. [25] examined their effectiveness for the understanding of specific information (typographic attributes bold and italic) by 30 sighted and 30 blind participants. A preliminary study of auditory rendition of typographical and/or punctuation information, using expressive speech synthesis, is presented in [26]. The aim is to increase the expressiveness of the already existing TtS system of France Telecom using prosodic rules. Four prosodic parameters are proposed for use: pitch, rate, volume, and break.

The above studies essentially propose rules for the implementation of the acoustic rendition of specific typographic attributes. It is obvious that a systematic methodology towards the acoustic rendition of typographic signals does not exist. The present work introduces the emotional-based mapping methodology for rendering font cues to auditory modality. The methodology is applied in a case study for font size and style. We determine the acoustic rendition of the font attributes by combining a text font-to-emotional state model and expressive speech synthesis. By conducting a number of psychoacoustic experiments, we determine the acoustic rendition of text font cues. Our ultimate goal is to incorporate automatic text font-to-speech mapping in DtA by emotional analogy between the visual (text font cues) and the acoustic (speech prosody) modalities.

In Section 2, we present a review on the relation of human emotions with typography and speech. In Section 3, first we present a preliminary study on direct mapping of typography based on the analysis of speech corpora. Then, based on the visual and acoustic modality emotional analogy, we introduce the emotion-based typography mapping. Following the proposed methodology, the emotional states are extracted and modeled on font style (plain, bold, italics, and bold-italics) and font size. The determination of the analogous prosodic cues (pitch, rate, and volume) was based on the model proposed by Schröder [27]. Then, each font cue is mapped into a value of a specific prosodic cue. As these values are below the human listener's discrimination level, we normalize them by applying linear quantization along with a psychoacoustic experiment in order to select the optimum font-to-speech devices. The final selected devices are evaluated in Section 4.

2 Human emotions, typography, and speech

Studies on emotions can be classified into i) categorical (discrete emotions) and ii) dimensional. The discrete emotion approach relies on a small set of emotions (e.g., the six basic emotions [28]: anger, disgust, fear, joy, sadness, and surprise). The number of the basic emotions differs among theorists. Plutchik [29] distinguished eight basic emotions: fear, anger, sorrow, joy, disgust, acceptance, anticipation, and surprise. Secondary ("non-basic" or "mixed") emotions are those that cannot be described solely by a basic emotion. For example, "hostility" can be defined as a mixture of "anger" and "disgust".

The dimensional theory [30] deals with emotions on the three dimensions of the emotional space, namely "Pleasure" (or "Valence"), "Arousal", and "Dominance" (or "Potency"). The dimension of "Pleasure" varies from negative to positive on the emotional poles and its middle represents a neutral affect. The dimension of "Arousal" varies from calm to highly aroused poles and the "Dominance" varies from controlled to in-control poles.

Discrete emotions can be mapped into the three-dimensional space of the emotional states. A well-known example is the Russell's circumplex [30]. The two dimensions of "Pleasure" and "Arousal" are represented on an XY grid, respectively. Another version of emotional grid is the Geneva Emotion Wheel [31].

2.1 Reader's emotional state modeling

Emotions can be incorporated in text documents either using the semantics of the content or through the visual typographic cues. Several studies focus on the semantics-based extraction and modeling of emotions from the content of the documents (i.e., [32–34]).

Document structure affects the reading comprehension, browsing, and perceived control [35]. Hall and Hanna [36] examine the effect of web page text/background color combination on readability, retention, aesthetics, and behavioral intention. Ethier et al. [37] studied the impact of four websites' interface features on the cognitive process that trigger online shoppers' emotions. Focusing on the typographic attributes and using the dimensional theory of emotions, Laarni [38] investigated the effects of color, font type/style on the "Pleasure", "Arousal", and "Dominance" scales according to the users' preferences. Furthermore, he examined the impact of color on document aesthetics (e.g., combinations of red font on green background were rated as the most unpleasant and black on white were considered the least arousing). Ho [39] in a review study on typography and emotions, concluded that most fonts and typefaces have a certain level of emotional potency. According to

the experimental study of Koch [40], participants responded to typefaces with statistically significant levels of emotion. Ohene-Djan et al. [41, 42] studied how the text's typographic elements can be used to convey emotions in subtitles mainly for the deaf and hearing-impaired people. They use font color and font size along with the emotions happily, sadly, sarcastically, excitedly, comically, fearfully, pleadingly, questioningly, authoritatively, and angrily. Using the TextTone system [43], emotions (e.g., happy, upset, disappointed, angry, very angry, ajd shocked) can be conveyed during online textual communication. This has been implemented by changing the typographic attributes. Moreover, Yannicopoulou [44] has examined the visual metaphor of emotions through the voice volume and letter size analogy. Based on the dimensional theory of emotions, a recent study [8] investigates how the typographic elements, like font style (bold, italics, and bold-italics) and font (type, size, color, and background color), affect the reader's emotional states "Pleasure", "Arousal", and "Dominance" (PAD). Finally, the preliminary quantitative results of a regression model [17]: a) revealed the impact of font/background color brightness differences on readers' emotional PAD space and b) showed that font type affects the "Arousal" and "Dominance" dimensions.

2.2 Expressive speech synthesis

Expressive Speech is "the speech which gives us information, other than the plain message, about the speaker and triggers a response to the listener" [45]. Emotion is seen as a type of expression, thus, expressive speech synthesis (ESS) is a method for conveying emotions (and other paralinguistic information) through speech, using the variations and differences of speech characteristics. There is a plethora of studies towards the development of ESS [46–48]. Many of them focus on the expression of specific emotions that can be extracted from the speech [46, 49, 50]. The term "variety of styles" has been introduced as a domain-dependent point of study. For example, ESS can convey messages such as "good-bad news", "yes-no questions" [51], "storytelling" [52], and "military" [53].

EES can be implemented by applying formant, waveform concatenation (mainly diphone-based), unit selection, or explicit prosody control synthesis [46, 47, 54]. In formant synthesis, the resulting speech synthesis is relatively unnatural, compared to concatenation based systems [47]. The most natural speech synthesis technique is unit selection or large database-based synthesis. Moreover, newer methodologies [48] optimize the existing ones or propose novel approaches such as expressivity-based selection of units, unit selection, and signal modification, as well as statistical parametric synthesis based on Hidden Markov Models [55].

2.3 Expressive speech synthesis: the dimensional approach

Schröder [27] developed a model for the elaboration of an ESS system using the dimensional approach of emotions, namely "Pleasure", "Arousal", and "Dominance" (PAD). The advantage of using this method is that the values in PAD dimensions are continuous. PAD values can be mapped in a specific emotional state (or variations of the emotion). For example, the emotions "happy/sad" can have variations like "quite happy/sad", "very happy/sad", and "less happy/sad". This model has been implemented and tested using the MARY TtS system [56]. Several equations describe how the prosodic parameters vary while changing the emotional states [27]. The parameters are distinguished as: i) "Standard" global parameters: pitch, range, speech rate, and volume, ii) "Non-standard" global parameters: pitch-dynamics and range-dynamics, and iii) specific entities like "GToBI accents" and "GToBI boundaries" (German Tones and Break Indices [GToBI]). Both values of the dependent (prosodic parameters) and the independent (emotional states) variables are continuous. Equation (1) presents Schröder's [27] general model:

$$S = F * E + I \tag{1}$$

where

$$S = \begin{bmatrix} S_1 \\ S_2 \\ \dots \\ S_n \end{bmatrix} \quad F = \begin{bmatrix} a_1^P & a_1^A & a_1^D \\ a_2^P & a_2^A & a_2^D \\ \dots & \dots & \dots \\ a_n^P & a_n^A & a_n^D \end{bmatrix} \quad I = \begin{bmatrix} I_1 \\ I_2 \\ \dots \\ I_n \end{bmatrix} \quad E = \begin{bmatrix} P \\ A \\ D \end{bmatrix},$$

where

S is the speech (prosodic) characteristics matrix; P is pleasure in [−100, 100]; A is arousal in [−100, 100]; D is dominance in [−100, 100]; F is the factors matrix; and I is the intercept (offset) matrix.

In the current study, we use the three basic prosodic parameters *pitch*, *rate*, and *volume*. According to Schröder's model, the way these parameters vary is described by the following equation:

$$\begin{bmatrix} \text{Pitch} \\ \text{Rate} \\ \text{Volume} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.3 & -0.1 \\ 0.2 & 0.5 & 0 \\ 0 & 0.33 & 0 \end{bmatrix} \cdot \begin{bmatrix} P \\ A \\ D \end{bmatrix}. \tag{2}$$

2.4 Speech and emotions across cultures and languages

It is believed that there is a certain universal cross-linguistic particularity behind emotions [57]. Scherer et al. [58] found that judges in nine countries, speaking different languages, "can infer four different emotions

and neutral state from vocal portrayals using content-free speech, with a degree of accuracy that is much better than chance". Also, "differences in hit rates across emotions and the error patterns in the confusion matrices are highly similar across all countries". There is evidence that the linguistic and/or the cultural similarity could play a role in vocal emotion recognition. Extending this work, Pell et al. [59] investigate emotion recognition across four languages (English, German, Hindi, and Arabic) using six distinct emotions. According to their results, the vocal emotion recognition is "largely unaffected by language or linguistic similarity". Moreover, Schröder [27], in his study of the emotional corpus analysis as a starting point for the synthesis rules, supports the hypothesis that "vocal emotion expression is very similar across languages". Based on a multilingual comparison study, Burkhardt et al. [60] reported on the effects of prosodic changes on emotional speech. According to their results, listeners in France, Germany, Greece, and Turkey correctly interpret semantically identical sentences expressing emotional-relevant content (neutral, joyful, friendly, threatening, frightened, and sad phrases). They found differences between the countries, and they could not estimate whether all effects were based on cultural difference alone. Thus, they claim that a cross-cultural global emotion simulation will not work as expected, and their findings indicate that results based on data-analysis of different cultures cannot be applied without reservations. This assumption cannot be totally supported because the deviations may be based on other facts than just the cultural difference. For example, a) their results were based on stimuli that were different across cultures; b) possibly the differences are coming from the speech synthesizer itself (they use diphone-based synthesis with different speakers from each language, and thus the quality of the synthetic speech for each language most probably is not the same); c) the underlying prosody was based on different speakers; and d) the translation of the sentences may have resulted in different semantics implying a different kind of appropriate emotion.

3 Prosodic mapping of typography

3.1 Preliminary study on direct mapping

The primary goal of this preliminary study is to observe how speakers vocally express typesetting, and specifically, "bold" and "italics". First, we recorded and analyzed the participants' vocal renditions of typographic elements, and then, through a discussion/interview, they asked to describe why they have selected the specific vocal expressions to render the typographic elements.

The text document used for the acoustical rendition contained approximately 1000 words. It was constructed by combining paragraphs from a Greek textbook (secondary high school Biology). The body of the text was in plain "Arial", 13pt font size. "Bold" and "italics" typesetting were applied on specific parts of the text, namely:

- Four words (two words in "bold" and two words in "italics")
- Four phrases (two sentences in "bold" and two sentences in "italics" with lengths of three words, respectively)
- Four sentences (one sentence in "bold" and one sentence in "italics" with lengths of five words, respectively, and a second set of two sentences with lengths of seven words, respectively)

Four participants (mean age = 35.5 years, SD = 3.4) were used for the creation of the speech corpus: one male and one female with a BSc degree in Physics and one male and one female with a BSc degree in Computer Science. All of them are experienced instructors in secondary school education, without any articulation problem.

The document's vocalization process was realized in the recording studio of the Department of Informatics and Telecommunication, University of Athens. Each participant was recorded separately. Before the recording session, each one was familiarized with the purpose of the study. They were introduced with the document. They were instructed to read it clearly and naturally with a stable tempo and volume. They could stop the entire process as they wished for a relaxing break or if they wanted to repeat the recording. Also, the supervisor of the process could interfere and ask the participant to repeat the recording in the case it did not meet the desired quality. During the process, when the participant "faced" a word, phrase, or sentence in typesetting, they had to pronounce it as she/he desired changing her/his pitch (low or high voice) and rate (faster or slower).

After the recording session, a prosody analysis of pitch and rate was performed using PRAAT [61]. The mean pitch baseline is dependent on the participant. It ranges from approximately 110 Hz (for a male participant) to 210 Hz (for a female participant). The word rate baseline had small deviation across participants ranging from 124 to 130 wpm.

The differences in pitch and word rate have similar behavior across participants. For the "bold" typesetting, the pitch differences were ranging from small (e.g., −7 %, for a female participant) to large (e.g., −36 %, for a male participant). The percentage word rate decrease had similar behavior (e.g., a female participant in a sentence case −23 %, and a male participant in a phrase case −56 %). Similar results were observed for the rendition of the "italics" typesetting, e.g., +11 and +49% for pitch differences and +14 and +48 % for word rate differences.

Then, participants were asked to describe why they have selected the specific manner they used to render "bold" and "italics" using pitch and speech rate variations. All of them answered that they rendered

- "bold" with a decreased pitch and speech rate, because they believe that "bold" expresses "emphasis" or "something important" and
- "italics" with increased pitch and speech rate, because it expresses something "that is less important", "a note", or "an explanation".

3.2 Acoustic rendition of typography using emotions: methodology

The primary goal of using typographic attributes in text documents is to distinguish parts of the text and to create a well-formed presentation of the content in order to augment the reading performance, attract the reader, and render semantics through the visual channel. We introduce the notion "typographic profile" of emotions, in a similar way to the "prosodic profile" [45]. The way typographic attributes are used in a document constitute its typographic profile. The profile defines the space within which each emotion is located (and vice-versa). Fig. 1 presents the prosodic and typographic profile, respectively.

Thus, emotions/emotional states constitute a bidirectional relation between the two modalities. The yellow arrow in Fig. 1 represents how we use the two profiles in this study in order to render the font cues into the acoustic modality. Our syllogism is as follows:

Typography induced emotions or emotional states can be described by (3):

$$E = F * T + B \tag{3}$$

where T is the typographic cue ($m \times 1$); E is the emotional state ($n \times 1$); F is the factors matrix ($n \times \mathrm{m}$); B is the offset matrix ($n \times 1$); m is the number of emotions; and E is a matrix that describes either the differences of the discrete emotions or the emotional states:

$$E = E_{\text{final}} - E_{\text{baseline}},$$

where E_{baseline} constitutes the emotional state for the corresponding typographic baseline and E_{final} the emotional state for the corresponding typographic attribute. We define as **typographic baseline** the most frequent value of the typographic attribute that appears in the whole document.

T is the matrix that describes the typographic cue. It can have the form of either a single variable polynomial equation

$$T = a_n \cdot t^n + \cdots + a_2 \cdot t^2 + a_1 \cdot t + b$$

or a multiple variable polynomial equation

$$T = a_n \cdot t_n + \cdots + a_2 \cdot t_2 + a_1 \cdot t_1 + b$$

In the case the emotions/emotional states are discrete over typography, then the correlation matrix is $F = 0$, and the emotion state variations equals to the offset matrix B.

The emotional variations affect the prosodic characteristics of the speech according to (4):

$$S = F' * E + B', \tag{4}$$

where S are the prosodic values ($m \times 1$); E is the emotional state ($n \times 1$); F is the factors matrix ($m \times \mathrm{n}$); and B is the offset matrix ($m \times 1$).

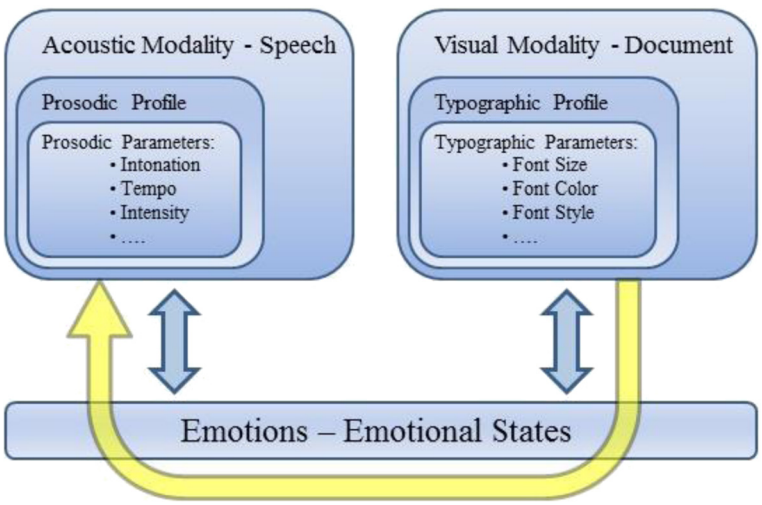

Fig. 1 The prosodic and typographic profiles

By combining (3) and (4)

$$\begin{aligned} S &= F' * (F * T + B) + B' \\ &= F' * F * T + F' * B + B' \end{aligned}$$

or

$$S = C * T + D, \quad (5)$$

where S are the prosodic elements (m × 1); T are the typographic cues ($m \times 1$); $C = F' * F$ is the factors matrix ($m \times m$); and $D = F' * B + B'$ is the offset matrix ($m \times 1$).

The above Eq. (5) essentially provides a mathematical description of the proposed model on the emotion-based rendering of typographic cues to the auditory modality.

Based on the abovementioned methodology, in the following sections, we describe the prosodic mapping of text font based on the dimensional theory of emotions in the case of style (bold, italics, and bold-italics) and font size.

3.3 Emotional state on font size and style

In order to study the emotional state elicited by the font size and style, we have selected the dimensional approach along with the Self-Assessment Manikin test. A computer-based experiment was designed, implemented and conducted [8] according to the International Affective Picture System guidelines [62]. The purpose of this experimental procedure was to assess the emotional states of the participants after reading a short text presented on the screen in various typographic cues.

Thirty native Greek undergraduate or postgraduate students participated, 16 male and 14 female, aging from 18 to 33 (Mean Age = 25.5, SD = 3.4), with normal color vision and normal or corrected-to-normal visual acuity. A 17-in LCD display with a resolution of 1024 × 768 was used to display the stimuli in a random order. The stimuli were created after applying the typographic cues (ten font sizes, four-font style × two-font type combinations) on an emotionally neutral Greek sentence of 46 words (Fig. 2). The sentence used in stimuli was determined as emotionally neutral in a previous pilot study [63]. The typographic cues were applied on the entire sentence. Each stimulus was presented to the participants for 15 s. The participants were asked to complete the 9-point PAD scale questionnaire, using the manikins provided by the SAM test.

Σε αυτή την έρευνα προσπαθούμε να αξιολογήσουμε πώς το μέγεθος των γραμμάτων ενός κειμένου, το είδος της γραμματοσειράς και ο τρόπος γραφής (απλή, πλάγια, έντονη, έντονη-πλάγια) που χρησιμοποιείται στο κείμενο και οι χρωματικοί συνδυασμοί του κειμένου και του υπόβαθρού του επηρεάζουν τα συναισθήματα του αναγνώστη.

Fig. 2 Stimulus sentence with font type "*Arial*" and font style "*italics*"

Table 1 presents the *font sizes* applied to create the stimuli and the mean values (rounded to the second decimal) of the participants' answers on each size along with their standard errors.

After applying a polynomial fitting on the mean values, the following set of equations was derived:

$$\begin{bmatrix} P \\ A \\ D \end{bmatrix} = \begin{bmatrix} 0.24942 & -0.00523 & 0 \\ -0.59264 & 0.02866 & -0.000427111 \\ -0.41265 & 0.01631 & -0.000199863 \end{bmatrix} * \begin{bmatrix} s \\ s^2 \\ s^3 \end{bmatrix} + \begin{bmatrix} -2.45287 \\ 3.58459 \\ 2.85307 \end{bmatrix}, \quad (6)$$

where P is Pleasure in [−1, 1]/adjusted $R^2 = 0.91$; A is Arousal in [−1, 1]/adjusted $R^2 = 0.78$; D is Dominance in [−1, 1]/adjusted $R^2 = 0.85$; and s is font size in px.

For the font style, plain, bold, italics, and bold-italics in both *Times New Roman* and *Arial*, with *black font color* on *white background* and *16 px font size* were used. The results revealed that only the font style and the interaction between the font style and font type is statistically significant in the "Pleasure" dimension. Also, "Arousal" and "Dominance" are not affected by the font type/style or their interaction (two-way repeated measures ANOVA, $p < 0.05$).

While using plain text on Arial, "Pleasure" is 0.13 (SE = 0.05)***, using bold is 0.10 (SE = 0.05)†, italics is 0.33 (SE = 0.07)*, and bold-italics is 0.26 (SE = 0.08)**. For Times New Roman, using plain text, "Pleasure" is 0.36 (SE = 0.06)*,

Table 1 Mean values of "Pleasure", "Arousal", and "Dominance" in the [−1,1] scale along with their standard errors, for ten different font sizes using Times New Roman font type, with black font color on white background

Font size (px)	Mean (SD)		
	Pleasure***	Arousal***	Dominance***
10	−0.47 (0.08)*	0.08 (0.12)	0.06 (0.12)
11	−0.30 (0.07)*	−0.05 (0.10)	0.04 (0.11)
12	−0.27 (0.07)*	−0.06 (0.10)	−0.07 (0.11)
13	−0.23 (0.08)*	−0.18 (0.10)^{a}	−0.10 (0.11)
14	−0.01 (0.05)	−0.36 (0.09)*	−0.21 (0.11)^{a}
15	0.13 (0.06)**	−0.37 (0.09)*	−0.40 (0.12)*
16	0.36 (0.06)*	−0.27 (0.10)*	−0.43 (0.11)*
18	0.28 (0.08)*	−0.21 (0.11)^{a}	−0.53 (0.11)*
26	0.43 (0.07)*	0.02 (0.10)	−0.31 (0.12)**
32	0.20 (0.07)*	−0.03 (0.10)	−0.22 (0.12)^{a}

Significantly different from 0.00 at $p < 0.01$*, $p < 0.05$**, and $p < 0.1^{a}$

*** One-way repeated measures ANOVA ($p < 0.05$) revealed that font size affects "Pleasure", "Arousal", and "Dominance"

using bold is 0.05 (SE = 0.06), italics is 0.17 (SE = 0.06)***, and bold-italics is 0.15 (SE = 0.07)*** (note: significantly different from 0.00 at $p < 0.001^{*}$, $p < 0.01^{**}$, $p < 0.05^{***}$, and $p < 0.1^{\dagger}$).

While font size increases (up to 26 px), the value of the "Pleasure" dimension is also increased. For greater font sizes, the "Pleasure" is decreased. On the other side, "Arousal" and "Dominance" decrease, up to 15 and 18 px, respectively, and then they start to increase. Moreover, the font type "Times New Roman" on plain text is more pleasant than the corresponding "Arial". All the other styles (bold, italics, and bold-italics) are more pleasant in the case of "Arial" than in "Times New Roman". Based on the abovementioned findings, we can say that in some cases of font size and style, the induced emotions are small; but in some other cases, we observe rather large emotional variations.

3.4 Modeling font size and type to expressive speech synthesis

After a statistical survey conducted on a large number of text books and newspapers [63], we found that the most frequent font sizes that define the baseline of the documents are 10, 12, and 14pt. Moreover, we observed that the increase of font size (e.g., in the titles or headlines) can be 12, 14, 16, 18, 20, and 22pt and the decrease (e.g., in footnotes) 8, 9, 10, 11, 12, and 13pt, depending on the font size baseline.

The physical length of 72 points font size is 2.54 cm (1 in.). Thus, one point equals to $d_{\mathrm{po}} = 1/72{*}2.54 = 0.035$ cm. For example, a font size of 10 px having 0.33 cm physical height, corresponds to 0.33 cm/$d_{\mathrm{po}} = 9.43{\sim}9$ points. In a similar way, 18 px having 0.594 cm physical height corresponds to 0.594 cm/$d_{\mathrm{po}} = 16.97{\sim}17$ points. One can use the following conversion equation from points to pixels (font size range 8 to 22pt):

$$[\text{font size in pixels}] = [\text{font size in points}] + 1 \tag{7}$$

Using equation (3) and the font size conversion Eq. (7), we can calculate the differences (in the [−100, 100] scale) of "Pleasure", "Arousal", and "Dominance" according to a number of font size variations from the baseline. The calculated values are presented in Table 2.

The differences from the plain text after applying bold, italics, and bold-italics on both Arial and Times New Roman can be described using Eq. (8) for the PAD dimensions

$$E_n = B_n, \tag{8}$$

where

Table 2 Emotional states' differences (in the [−100,100] scale) resulted from the font size variations from the baseline

Font size increment/ decrement (pt)		Emotional state		
		Pleasure	Arousal	Dominance
+2pt	*10 → 12*	24.78	−17.95	−21.55
+2pt	*12 → 14*	20.60	−8.35	−14.74
+2pt	*14 → 16*	16.41	−0.79	−8.88
+4pt	*10 → 14*	45.38	−26.29	−36.29
+4pt	*12 → 16*	37.01	−9.14	−23.62
+4pt	*14 → 18*	28.64	3.91	−12.88
+6pt	*10 → 16*	61.79	−27.09	−45.17
+6pt	*12 → 18*	49.24	−4.43	−27.61
+6pt	*14 → 20*	36.68	12.07	−12.93
+8pt	*10 → 18*	74.02	−22.38	−49.16
+8pt	*12 → 20*	57.28	3.73	−27.67
+8pt	*14 → 22*	40.54	21.64	−10.02
−1pt	*10 → 9*	−13.96	13.22	13.63
−1pt	*12 → 11*	−11.87	7.65	9.86
−1pt	*14 → 13*	−9.77	3.10	6.58
−2pt	*10 → 8*	−28.96	29.60	29.32
−2pt	*12 → 10*	−24.78	17.95	21.55
−2pt	*14 → 12*	−20.60	8.35	14.74

$E_n = \begin{bmatrix} P_n \\ A_n \\ D_n \end{bmatrix}$, where n: p (plain) or b (bold) or i (italics) or bi (bold-italics).

In the case of Arial,

$$B_p = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, B_b = \begin{bmatrix} -3 \\ 0 \\ 0 \end{bmatrix}, B_i = \begin{bmatrix} 20 \\ 0 \\ 0 \end{bmatrix}, B_{bi} = \begin{bmatrix} 13 \\ 0 \\ 0 \end{bmatrix}.$$

In the case of Times New Roman,

$$B_p = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, B_b = \begin{bmatrix} -31 \\ 0 \\ 0 \end{bmatrix}, B_i = \begin{bmatrix} -19 \\ 0 \\ 0 \end{bmatrix}, B_{bi} = \begin{bmatrix} -21 \\ 0 \\ 0 \end{bmatrix}.$$

The mapping of typography into speech can be accomplished by applying the above outcomes to the general Eq. (5) of the proposed model using XSLT files in the OpenMary System [56]. The results are presented in Tables 3 and 4 for the cases of font size and type. The corresponding prosodic values of the typographic

Table 3 Pitch, rate, and volume differences (absolute and percentage values) on font size using three different baselines

Font sizes		Prosodic differences					
		Pitch		Rate		Volume	
		Hz	%	wpm	%	0–100 scale	%
+2pt	*10 → 12*	−0.75	−0.68	−7.24	−4.02	44.08	−5.92
+2pt	*12 → 14*	1.03	0.94	−0.09	−0.05	47.25	−2.75
+2pt	*14 → 16*	2.29	2.08	5.20	2.89	49.74	−0.26
+4pt	*10 → 14*	0.28	0.25	−7.33	−4.07	41.32	−8.68
+4pt	*12 → 16*	3.32	3.02	5.09	2.83	46.98	−3.02
+4pt	*14 → 18*	5.33	4.85	13.82	7.68	51.29	1.29
+6pt	*10 → 16*	2.57	2.34	−2.14	−1.19	41.06	−8.94
+6pt	*12 → 18*	6.36	5.78	13.73	7.63	48.54	−1.46
+6pt	*14 → 20*	8.58	7.80	24.07	13.37	53.98	3.98
+8pt	*10 → 18*	5.60	5.09	6.50	3.61	42.61	−7.39
+8pt	*12 → 20*	9.61	8.74	23.98	13.32	51.23	1.23
+8pt	*14 → 22*	11.55	10.50	34.07	18.93	57.14	7.14
−1pt	*10 → 9*	1.21	1.10	6.88	3.82	54.36	4.36
−1pt	*12 → 11*	0.12	0.11	2.61	1.45	52.52	2.52
−1pt	*14 → 13*	−0.71	−0.65	−0.72	−0.40	51.02	1.02
−2pt	*10 → 8*	3.05	2.77	16.22	9.01	59.77	9.77
−2pt	*12 → 10*	0.75	0.68	7.24	4.02	55.92	5.92
−2pt	*14 → 12*	−1.03	−0.94	0.09	0.05	52.75	2.75

baseline are pitch = 110 Hz, rate = 180 words/min. The volume is given in a 0–100 scale with a baseline of 50.

The prosodic variations in the abovementioned tables have small variations that cannot be discriminated by the human listener (more details in Section 3.5.1). In order to augment these variations, we propose the normalization of the values using linear quantization and the selection of the optimum prosodic variations.

Table 4 Pitch, rate, and volume differences (absolute and percentage values) on font type using plain text as baseline

		Bold	Italics	Bold-italics
Times New Roman				
Pitch	Hz	−3.1	−1.9	−2.1
	%	−2.8	−1.7	−1.9
Rate	wpm	11.2	−6.8	−7.6
	%	−6.2	−3.8	−4.2
Volume	0–100 scale	0.0	0.0	0.0
	%	0.0	0.0	0.0
Arial				
Pitch	Hz	−0.3	2.0	1.3
	%	−0.3	1.8	1.2
Rate	wpm	−1.1	7.2	4.7
	%	−0.6	4.0	2.6
Volume	0–100 scale	0.0	0.0	0.0
	%	0.0	0.0	0.0

3.5 Prosody normalization

3.5.1 Minimum perceived prosodic differences

3.5.1.1 Pitch Pitch differences of more than three semitones can be discriminated reliably [64]. But, pitch differences of 1.5 semitones create reliable differences in the perception of prominence [65]. Using a male voice with pitch F0 = 120 Hz, the speech stimuli discriminability is a) 1 Hz with flat F0 contours and b) 2 Hz with rising F0 contours, respectively [66, 67]. Truillet et al. [18] propose a 13 % increment of the default pitch. Xydas et al. [13], using a male voice with F0 = 110 Hz as a baseline, proposed a 15 % increment of the default pitch. With a two-semitone difference (~12 %), neighboring pitch components may be resolvable even when they have a similar amplitude [68].

3.5.1.2 Speech rate Xydas et al. [13] proposed a 15 % of rate variation (using as baseline 140 wpm). For natural speech rates in the range 125–225 wpm, the increase of the speech rate is not significantly associated with decreasing recall for native listeners [69]. Exceeding the boundary of 225 wpm, there is an accelerating decline in comprehension by native listeners.

3.5.1.3 Volume It is experimentally revealed that the "linear increase of speech signal amplitude by 1.3 on stressed syllables is well assessed by listeners evaluating the resulting synthesized speech" (increase of intensity level by 1–3 dB) [70].

3.5.1.4 Pauses High-quality speech without grammatical speech pauses within sentences can be highly intelligible and acceptable. But, as soon as the speech quality is less than normal, or the speech is listened in noisy conditions, the introduction of grammatical speech pauses can help to maintain intelligibility [64]. In general, "it can be observed that the contributions of prosody to speech perception become more important when the segmental quality of speech or the listening conditions become less favorable" [64]. Pauses can be considered as "markers of information structure" [71, 72]. Intonational phrases (clearly signaled by silence, as well as by the characteristic pitch movement) are required between utterances and at punctuation boundaries. Phonological phrases (signaled by silence rather by the characteristic pitch movement only) "may be harder to place with certainty and to evaluate" [72]. Voices with a high speaking rate combined with long utterance breaks seem to be preferable to listeners [73].

3.5.2 Quantization of the results

The levels of linear quantization are defined in Tables 5 and 6 (for the font size and the font type, respectively) by calculating the difference between the minimum and the maximum values for each prosodic parameter and dividing by the six levels of change. The number of levels is considered as the optimum and is derived from observations while trying to get the minimum duplicate cases of the prosodic variations.

By applying the above linear quantization to the results in Tables 3 and 4, we obtain the levels of prosodic variation normalization for both the font size and font style (Tables 7 and 8).

The optimum values for each level of the prosodic parameters used in this work are:

- 12 % for pitch (corresponds approximately to two semitones),
- 10 % or 14 wpm for speech rate, and
- 15 % for intensity corresponding to more than ±1 dB change.

Using ±6 levels, the minimum/maximum boundaries for the prosodic parameters, are respectively, pitch 96.8/189.2 Hz, rate 112/210 wpm, and volume −5.19/+4.08 dB.

Tables 9 and 10 present the normalized results for font size and font style, respectively, after applying the six-level linear quantization and the optimum prosodic values. The prosodic baseline is: Pitch = 110 Hz, Rate = 140 wpm, and Volume = 0 dB.

Following the suggestion of Huang [72], before and after each prosodic variation, we used a speech pause of 150 ms, as utterance breaks of such duration seem to be preferable to listeners.

Table 5 The six-level linear quantization of the prosodic cues on font size

	Pitch (Hz)	Rate (%)	Volume (%)
Minimum values	−1.03	−4.07	−8.94
Maximum values	11.55	18.93	9.77
Peak to peak values	12.58	23.00	18.71
Levels of variation		Limits for each level	
±1	±2.10	±3.83	±3.12
±2	±4.20	±7.66	±6.24
±3	±6.30	±11.49	±9.36
±4	±8.40	±15.32	±12.48
±5	±10.50	±19.15	±15.60
±6	±12.60	±22.98	±18.72

Table 6 The six-level linear quantization of the prosodic cues on font type

	Pitch (Hz)	Rate (%)
Minimum values	−3.1	−6.2
Maximum values	2.0	4.0
Peak to peak values	5.1	10.2
Levels of variation	Limits for each level	
±1	±1.00	±2.00
±2	±2.00	±4.00
±3	±3.00	±6.00
±4	±4.00	±8.00
±5	±5.00	±10.00
±6	±6.00	±12.00

3.6 Optimum acoustic rendition of text font

The model described in the previous sections proposes multiple cases for each typographic cue (three cases for each font size and two cases for each font style). Using many prosodic cues in speech and mapping the same font attribute in different ways can be confusing to the listener. The following experiment has been designed in order to select the optimum acoustic rendition of text font cues.

Twenty-eight undergraduate or postgraduate students, 18 male and 10 female, participated, aging from 19 to 33 (Mean Age = 27.7, SD = 3.2). Their native language is

Table 7 The levels of the prosodic cues on font size

Font size alterations		Variations		
		Pitch (level)	Rate (level)	Volume (level)
+2pt	*10 → 12*	−1	−2	−2
+2pt	*12 → 14*	+1	−1	−1
+2pt	*14 → 16*	+2	+1	−1
+4pt	*10 → 14*	+1	−2	−3
+4pt	*12 → 16*	+2	+1	−1
+4pt	*14 → 18*	+3	+3	+1
+6pt	*10 → 16*	+2	−1	−3
+6pt	*12 → 18*	+4	+2	−1
+6pt	*14 → 20*	+5	+4	+2
+8pt	*10 → 18*	+3	+1	−3
+8pt	*12 → 20*	+5	+4	+1
+8pt	*14 → 22*	+6	+5	+3
−1pt	*10 → 9*	+1	+1	+2
−1pt	*12 → 11*	+1	+1	+1
−1pt	*14 → 13*	−1	−1	+1
−2pt	*10 → 8*	+2	+3	+4
−2pt	*12 → 10*	+1	+2	+2
−2pt	*14 → 12*	−1	+1	+1

Table 8 The levels of the prosodic cues on font style

	Bold	Italics	Bold-italics
Times New Roman			
Pitch (level)	−4	−2	−3
Rate (level)	−4	−2	−3
Volume (level)	–	–	–
Arial			
Pitch (level)	−1	+3	+2
Rate (level)	−1	+3	+2
Volume (level)	–	–	–

Greek. All of them had no hearing problem and normal or corrected-to-normal visual acuity.

Twenty-four stimuli in the Greek language were used. Each stimulus was a short sentence where a typographic attribute was applied on a phrase in the sentence (Fig. 3). The sentences were in plain text and the phrases were in bold, italics, bold-italics, and plain text with font sizes 2, 4, 6, 8, −1, or −2pt. The sentences were converted into speech by implementing the rules described in Section 3.5.2 using the document-to-audio platform which is based on the DEMOSTHeNES TtS system [13]. Each stimulus was presented in both the visual (printed) and the acoustic modalities simultaneously. Each participant was familiarized with the experiment listening to a few examples.

Participants could hear each stimulus only two times through high-quality headphones (AKG K271 MKII). Then, they were asked if they agree that the phrase in each acoustic stimulus has a specific font attribute (e.g., bold). They had to choose an answer between a 5-scale response: 1 corresponds to "No", 3 to "I do not know", and 5 to "Yes". The intermediate answers correspond to "possibly No" and "possibly Yes", respectively. The mean time of the responses for all the stimuli was approximately 18 min.

t test was used to examine if the mean values of the answers are statistically greater than the value 3 (that corresponds to "I do not know"). Font style preferences are significantly different ($p < 0.05$) in the cases of: "bold" (Mean = 3.50 SD = 1.48) and "italics" on Arial (Mean = 3.54, SD = 1.40), as well as for "bold-italics" on Times New Roman (Mean = 3.57, SD = 1.50). In the case of font size ($p < 0.05$), in one implementation of 2pt increase Mean = 3.93, SD = 1.02, and in one implementation of 1pt decrease, Mean = 3.46, SD = 1.23. Table 11 summarizes the statistically significant results.

Participants selected small decrement of pitch and rate for the rendition of "bold" and large decrement for "bold-italics". Moreover, "italics" were rendered with a large increment in pitch and rate. Comparing the results with those presented in Section 3.1 (direct acoustic rendition of typography using corpus analysis), we observe that the acoustic mapping of "bold" and "italics" have

Table 9 The normalized results of the prosodic cues for the font size variations

Font sizes		Percentage differences			Absolute values		
		Pitch (%)	Rate (%)	Volume (%)	Pitch (Hz)	Rate (wpm)	Volume (dB)
+2pt	*10 → 12*	−12	−20	−30	96.8	112	−3.10
+2pt	*12 → 14*	12	−10	−15	123.2	126	−1.41
+2pt	*14 → 16*	24	10	−15	136.4	154	−1.41
+4pt	*10 → 14*	12	−20	−45	123.2	112	−5.19
+4pt	*12 → 16*	24	10	−15	136.4	154	−1.41
+4pt	*14 → 18*	36	30	15	149.6	182	1.21
+6pt	*10 → 16*	24	−10	−45	136.4	126	−5.19
+6pt	*12 → 18*	48	20	−15	162.8	168	−1.41
+6pt	*14 → 20*	60	40	30	176.0	196	2.28
+8pt	*10 → 18*	36	10	−45	149.6	154	−5.19
+8pt	*12 → 20*	60	40	15	176.0	196	1.21
+8pt	*14 → 22*	72	50	45	189.2	210	3.23
−1pt	*10 → 9*	12	10	30	123.2	154	2.28
−1pt	*12 → 11*	12	10	15	123.2	154	1.21
−1pt	*14 → 13*	−12	−10	15	96.8	126	1.21
−2pt	*10 → 8*	24	30	60	136.4	182	4.08
−2pt	*12 → 10*	12	20	30	123.2	168	2.28
−2pt	*14 → 12*	−12	10	15	96.8	154	1.21

Table 10 The normalized results of the prosodic cues for the font style

	Bold	Italics	Bold-italics		Bold	Italics	Bold-italics
Times New Roman							
Pitch (%)	−48	−24	−36	pitch (Hz)	57	84	70
Rate (%)	−40	−20	−30	rate (wpm)	84	112	98
volume (%)	–	–	–	volume (dB)	–	–	–
Arial							
Pitch (%)	−12	36	24	pitch (Hz)	97	150	136
Rate (%)	−10	30	20	rate (wpm)	126	182	168
Volume (%)	–	–	–	volume (dB)	–	–	–

similar behavior in both studies. In the case of smaller font sizes, the acoustic mapping can be expressed by a small increment of pitch and a decrement of speech rate and volume. For the case of large font sizes, the acoustic mapping can be expressed by an increment of all the three prosodic parameters (pitch, rate, and volume).

4 Evaluation of the model

In order to evaluate the proposed model, we conducted an additional experiment. We examine whether the listeners can recognize the font cues (bold, italics, or bold-italics, and the increase or decrease of font size) without any prior familiarization with the prosodic mapping approach.

Eleven males and eight females participated in the experiment, 20 to 32 years of age (Mean Age = 27.8 years of age, SD = 3.2). All of them were undergraduate or postgraduate students and naive to any previous experiment of this work; their native language is Greek.

In total, 20 Greek stimuli were used, 12 for the font styles (4 stimuli for each of the 3 styles bold, italics and bold-italics) and 8 for the font size (4 stimuli for increasing and 4 for decreasing the font size). Each stimulus was a sentence with approximately 5-s duration (Fig. 4). The sentences were converted into speech using the document-to-audio platform [13] and the prosodic model was applied to a phrase in the sentence or to the entire sentence.

There was no training session for the participants. Two sessions were designed for evaluating the acoustic renditions, the first for the font style and the second for the font size. Each acoustic stimulus was simultaneously presented with the corresponding visual version of the sentence in plain text. The participants were asked to identify which typographic attribute they believe is associated within the sentence they hear (possible answers for the font style: "bold, italics, bold-italics or none" and for the font size: "size increase, size decrease or none").

Οι λέξεις **δοκιμαστικό κείμενο** είναι σε έντονα γράμματα.

Fig. 3 A sample stimulus (*bold case*) for the optimum acoustic rendition of font type and style

The identification results are presented in Figs. 5 and 6, respectively. In order to examine statistically the differences between the four stimuli for each case of the font style and size, we grouped the results into "Success" and "No Success". As Success, we marked the correct answers and No success all the wrong answers. For example, in the case of the bold font style, as Success, we marked the answers that identified as bold and the answers italics, bold-italics, and none are grouped as No Success. The results of the non-parametric Cochran's Q test for each of the three-font style and two-font size cases are:

$$\mathrm{MPSR}_{\mathrm{bold}} = 47.4\ (15.5), \chi^2(3) = 6.240,\ p = 0.100$$

$$\mathrm{MPSR}_{\mathrm{italics}} = 56.6\ (2.7), \chi^2(3) = 0.200,\ p = 0.978$$

$$\mathrm{MPSR}_{\mathrm{bold\text{-}italics}} = 34.3\ (10.1), \chi^2(3) = 4.400, p = 0.221$$

$$\mathrm{MPSR}_{\mathrm{increment}} = 52.6\ (16.6), \chi^2(3) = 6.207,\ p = 0.102$$

Table 11 The statistically significant results with the levels of the increment/decrement and the corresponding percentage differences and their prosodic absolute value

	Bold	Italics	Bold-italics	Size +2	Size -1
Pitch (level)	−1	+3	−3	+1	+1
Rate (level)	−1	+3	−3	−1	+1
Volume (level)	–	–	–	−1	+2
Pitch (%)	−12	36	−36	12	12
Rate (%)	−10	30	−30	−10	10
Volume (%)	–	–	–	−15	30
Pitch (Hz)	97	150	70	123	123
Rate (wpm)	126	182	98	126	154
Volume (dB)	–	–	–	−1.41	2.28

Εικοσιτετράωρη πρόβα των αγροτών κόβει στα δύο την Ελλάδα.

Fig. 4 A sample stimulus for the evaluation process

$$\mathrm{MPSR}_{\mathrm{decrement}} = 56.6\ (2.7), \chi^2(3) = 0.273,\ p = 0.965$$

MPSR mean percentage success rate. Standard Deviation is given in parentheses after the value of MPSR.

Thus, Cochran's *Q* test did not indicate any differences among the four stimuli for each of the three-font style and two-font size cases ($p = 0.05$).

In general, participants recognized the font style and size. The font size is perceived better than the font style. In detail, italics were recognized successfully in all cases. Font sizes are also recognized with success. In case 1 of size increase, most answers (42.1 %) declared size decrement (the correct answers were marginally lower 36.8 %). Bold-italics failed to be recognized by the participants and were confused with bold (almost all cases were assessed by the majority of the participants as bold). This is probably due to the similar prosodic variations in both cases of bold and bold-italics (decrease of pitch and rate).

5 Conclusions

In this work, we proposed a novel, modular and language-independent mathematical approach for the acoustic rendition of the document's font, based on the emotional analogy between the visual typographic cues and the speech prosody at the acoustic level. The approach is based on the human emotions-emotional states as the medium for the acoustic rendition of text font by combining font-to-emotions and expressive speech synthesis models. The proposed methodology is examined in the case of specific font sizes and styles and has been psychoacoustically evaluated by listeners with promising recognition results.

Some major findings from the study are: a) bold can be successfully expressed with a 12 % pitch decrease and a 10 % speech rate decrease; b) italics can be successfully expressed by a 36 % pitch increase and 30 % rate increase; c) font size increase by +2pt can be expressed by a 12 % pitch increase, a 10 % speech rate decrease, and a 15 % volume decrease; d) font size decrease by −1pt can be expressed by a 12 % pitch increase, a 10 % rate increase, and a 30 % volume increase. The results denoted the lack of appropriate rendition of bold-italics. As the participants were not trained with the acoustic representation of the typography, it would be very interesting to examine, in a future study, the proposed methodology in an experiment with trained participants with the same prosodic variations in order to examine "if" and "how much" a training session possibly optimize listeners' performance. Focusing on "bold" and "italics", the results of the current study show the same behavior with the results observed in a preliminary corpus analysis, during the direct mapping of the specific typesetting, presented in Section 3.1. Thus, there are strong indications, that direct mapping and through emotional states have similar expressions of typesetting.

In a previous study, Truillet et al. [18] present the results for the acoustic rendition of "bold" typesetting on salient words, in order to investigate the acoustic rendition of typesetting effect on memory, using verbal description of the typographic attribute or increasing the default pitch by 13 %, on both sighted and blind listeners (direct mapping). They observed no significant differences between groups' memorization and between

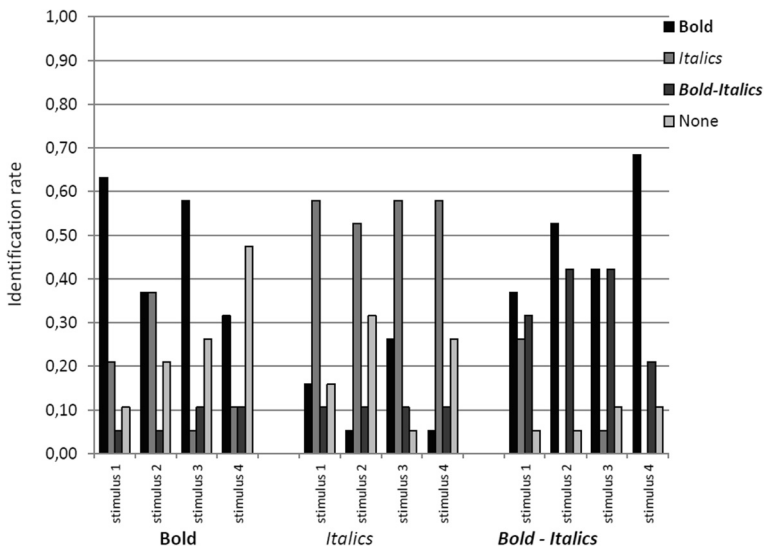

Fig. 5 The identification rate over the font style variations

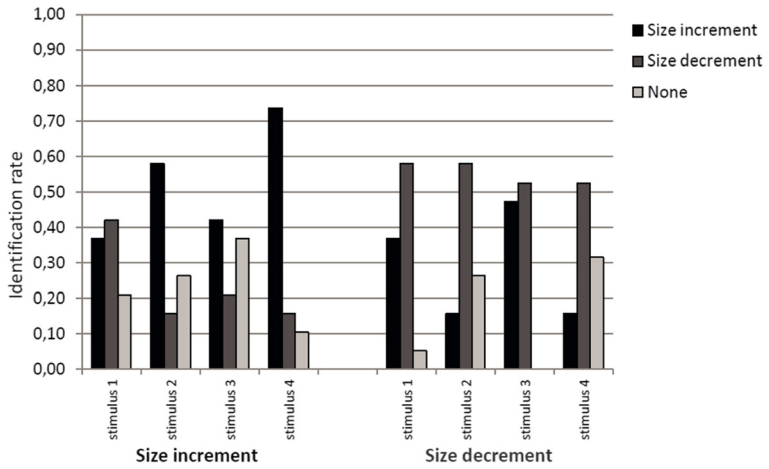

Fig. 6 The identification rate over the font size increase/decrease

different versions of salient words' expression. But, they present strong indications that the enriched version (verbalized description or prosodic differences while expressing the typographic attribute) has better performance than the neutral one, and the prosodic rendition has a better effect compared to the neutral and verbal description approach. Moreover, according to Asakawa et al. [21], the auditory interface of web pages, using two different ring tones to denote weak and strong emphasis level, is more useful for intuitive recognition. The emphasis level in the document was determined in [21] by the use of font size and style, giving priority to font size. For example, a part of text with much larger font sizes than the body of the main text is assigned to the strong emphasis and slightly larger size to weak emphasis level (indirect methodology using the semantics of typography of two different emphasis levels). Argyropoulos et al. [25] reported that blind student listeners have better performance under several conditions of typesetting acoustic rendition than their sighted peers. They used "bold" and "italics" and they assigned three different semantic conditions, "strong" for "bold", "definition" and "emphasis" for "italics". The three different conditions (along with an additional condition with no modifications) were acoustically rendered using the DEMOSTHeNES TtS system [13] and modifications of the pitch (main criterion), word rate, and speech volume. The way typographic cues were acoustically rendered was based on a literature survey and the experience of the research team. It is worth to note that in the case of sighted students, no condition was particularly more effective compared to any other condition and also the profile of the sighted group with regard to the prosody effects was "flat", or otherwise unaffected by the condition variable.

All the abovementioned studies do not follow a strict formalism for the acoustic rendition of typographic elements. Moreover, they study a small number of specific typographic cues (e.g., in [18] only the "bold" is examined), or they do not follow a consistent way on how they map typography into the acoustic modality. In the present study, we tried to overcome these limitations by introducing a mathematical description of the acoustic rendition of typography. Also, the direct mapping approaches mentioned above do not propose an optimum acoustic rendition of the typography. Thus, we cannot conclude that the direct mapping is better than the proposed indirect approach based on the emotional states.

In another future study, we intend to change the minimum perceived levels of pitch, rate and volume, in order to eliminate the pitfalls of the current study and to optimize listeners' performance (e.g., the misunderstanding of bold-italics and bold during their acoustic rendition) or the use of non-linear quantization. Also, the study of acoustic rendition of typographic attributes in a similar methodology using discrete emotions (instead of the dimensional approach) seems very interesting. This would extend the use of the current methodology in systems than can process a reader's discrete emotions derived from the typographic cues and the TtS systems that only support discrete emotion approach.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This research has been partially co-financed by the European Union (European Social Fund—ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF), Research Funding Project: THALIS-University of Macedonia, "KAIKOS: Audio and Tactile Access to Knowledge for Individuals with Visual Impairments" MIS 380442. The publication of this work has been funded by the National and Kapodistrian University of Athens.

Received: 13 July 2015 Accepted: 3 March 2016
Published online: 15 March 2016

References

1. M McLuhan, Q Fiore, *The Medium is the Message* (Gingko Press, Berkeley, California, 2005)
2. RF Lorch, Text-signaling devices and their effects on reading and memory processes. Educ. Psychol. Rev. **1**(3), 209–234 (1989)
3. JH Spyridakis, Signaling effects: a review of the research—part I. J. Tech. Writing. Commun. **19**(3), 227–240 (1989)
4. J Lemarié, H Eyrolle, JM Cellier, Visual signals in text comprehension: How to restore them when oralizing a text via a speech synthesis? Comput. Hum. Behav. **22**(6), 1096–1115 (2006)
5. R Lorch, HT Chen, J Lemarié, Communicating headings and preview sentences in text and speech. J. Exp. Psychol. Appl. **18**(3), 265–276 (2012)
6. ZH Han, ES Park, C Combs, Textual enhancement of input: issues and possibilities. Appl Linguist **29**(4), 597–618 (2008)
7. T van Leeuwen, Towards a semiotics of typography. Inform. Des. J. **14**(2), 139–155 (2006)
8. Tsonos, G Kouroupetroglou, Modeling reader's emotional state response on document's typographic elements. Adv. Hum. Comput. Interact. Article ID 206983, (2011). doi:10.1155/2011/2069832011
9. Print disabled (2016) https://en.wikipedia.org/wiki/Print_disability. Accessed 8 March 2016
10. J Blansett, Digital discrimination. Libr. J. **133**(13), 26–29 (2008)
11. S Narayanan, Text-to-speech synthesis. In by J Benesty, M Sondhi, Y Huang (Eds), *Handbook of Speech Processing*, (Springer, Berlin, Heidelberg, 2008), p. 411
12. D Freitas, G Kouroupetroglou, Speech technologies for blind and low vision persons. Technol. Disabil. **20**(2), 135–156 (2008)
13. G Xydas, V Argyropoulos, K Th, G Kouroupetroglou, An experimental approach in recognizing synthesized auditory components in a Non-visual interaction with documents, in *Proceedings of the 11th International Conference on Human-Computer Interaction (HCII '05)*, 2005, pp. 1–10
14. F Fourli-Kartsouni, K Slavakis, G Kouroupetroglou, S Theodoridis, A Bayesian network approach to semantic labelling of text formatting in XML corpora of documents, in *Universal Access in HCI, Part III, HCII 2007. Lecture Notes in Computer Science*, vol. 4556 (Springer, Verlag Berlin Heidelberg, 2007), pp. 299–308
15. G Xydas, G Kouroupetroglou, Text-to-speech scripting interface for appropriate vocalisation of e-texts, in *Proceedings of 7th European Conference on Speech Communication and Technology (EUROSPEECH 2001)*, 2001, pp. 2247–2250
16. G Xydas, D Spiliotopoulos, G Kouroupetroglou, Modelling emphatic events from non-speech aware documents in speech based user interfaces, in *Proceedings of the 10th International Conference on Human-Computer Interaction (HCII '03)*, vol. 2, 2003, pp. 806–810
17. D Tsonos, G Kouroupetroglou, D Deligiorgi, Regression modeling of reader's emotions induced by font based text signals, in AHCI/HCII 2013, part II, in *Lecture Notes in Computer Science*, ed. by C Stephanidis, M Antona, vol. 8010 (Springer, Verlag Berlin Heidelberg, 2013), pp. 434–443
18. P Truillet, B Oriola, JL Nespoulous, N Vigoroux, Effect of sound fonts in an aural presentation, in *Proceedings of the 6th ERCIM Workshop (UI4ALL 2000)*, 2000, pp. 135–144
19. K Kallinen, Using sounds to present and manage information in computers, in *Proceedings of the 2003 Informing Science and Information Technology Education Joint Conference* (Informing Science Institute, California, 2003), p. 1031
20. T Ph, B Oriola, N Vigouroux, Multimodal presentation as a solution to access a structured document, in *Proceedings of 6th World-Wide-Web Conference, Santa-Clara*, 1997
21. C Asakawa, H Takagi, S Ino, T Ifukube, Auditory and tactile interfaces for representing the visual effects on the Web, in ACM ASSETS 2002, Edinburgh, Scotland, UK, 8-10 July 2002
22. J Fackrell, H Vereecken, J Buhmann, JP Martens, BV Coile, Prosodic Variation with Text Type, in Proceedings of the 6th Int. Conf. on Spoken Language Processing (ICSLP 2000), vol.3, Beijing, China, October 2000, p. 231-234
23. W3C: CSS Speech Module (2015) http://www.w3.org/TR/css3-speech/. Accessed 8 March 2016
24. W3C: Aural Style Sheets (2015) http://www.w3.org/TR/CSS2/aural.html. Accessed 8 March 2016
25. V Argyropoulos, G Sideridis, G Kouroupetroglou, G Xydas, Auditory discriminations of typographic attributes of documents by students with blindness. Br. J. Vis. Impairment. **27**(3), 183–203 (2009)
26. J Launay, L Segalen, L Kanellos, T Moudenc, C Otesteanu, A David, G Fang, J Jin, Speech Expressiveness: Modeling and Implementing the Expressive Impact of Typographic and Punctuation Marks for Textual Inputs, in Proceedings of the 3rd International Conference on Information and Communication Technologies: From Theory to Applications, 2008, pp. 1–6
27. M Schröder, Expressing degree of activation in synthetic speech. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1128–1136 (2006)
28. P Ekman, An argument for basic emotions. Cogn. Emotion. **6**(3), 169–200 (1992)
29. R Plutchik, A general psychoevolutionary theory of emotion. Emotion. Theory. Res. Exp. **1**(3), 3–33 (1980)
30. KR Scherer, What are emotions? And how can they be measured? Soc. Sci. Inform. **44**(4), 693–727 (2005)
31. T Bänziger, V Tran, K Scherer, *The Geneva Emotion Wheel: a Tool for the Verbal Report of Emotional Reactions* (ISRE 2005, Bari, Italy, 2005)
32. AC Boucouvalas, Real time text-to-emotion engine for expressive internet communications, in emerging communication: studies on new technologies and practices in communication, in *Book Series IOS Press—Being There: Concepts, Effects and Measurement of User Presence in Synthetic Environments*, ed. by G Riva, F Davide, W IJsselsteijn, vol. 5, 2002, pp. 306–318
33. X Zhe, D John, AC Boucouvalas, Text-to-emotion engine: tests of user preferences, in *Proceedings of the IEEE International Symposium on Consumer Electronics (ISCE 2002)*, 2002, pp. B25–B30
34. S Owsley, S Sood, KJ Hammond, Domain specific affective classification of documents, in *Proceedings of the AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006, pp. 181–183
35. F Calisir, M Eryazici, MR Lehto, The effects of text structure and prior knowledge of the learner on computer-based learning. Comput. Hum. Behav. **24**(2), 439–450 (2008)
36. RH Hall, P Hanna, The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. Behav. Inform. Technol. **23**(3), 183–195 (2004)
37. J Ethier, P Hadaya, J Talbot, J Cadieux, Interface design and emotions experienced on B2C Web sites: empirical testing of a research model. Comput. Hum. Behav. **24**(6), 2771–2791 (2008)
38. J Laarni, Effects of color, font type and font style on user preferences, in *Adjunct Proceedings of HCI International 2003*, ed. by C Stephanidis (Crete University Press, Heraklion, 2003), p. 31
39. AG Ho, *Typography Today: Emotion Recognition in Typography. 5th International Association of Societies of Design Research*, vol. 1+2 (Japanese Society for the Science of Design, Tokyo. Japan, 2013), pp. 5573–5582
40. BE Koch, Emotion in typographic design: an empirical examination. Visible Lang. **46**(3), 206–227 (2012)
41. J Ohene-Djan, J Wright, K Combie-Smith, Emotional subtitles: a system and potential applications for deaf and hearing impaired people, in *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision and Hearing Impairments: Assistive Technology for All Ages*, 2007
42. J Ohene-Djan, R Shipsey, E- subtitles: emotional subtitles as a technology to assist the deaf and hearing-impaired when learning from television and film, in *Proceedings of the 6th International Conference on Advanced Learning Technologies*, 2006, pp. 464–466
43. A Kalra, K Karahalios, TextTone: expressing emotion through text. Lect. Notes Comput. Sci. **3885**, 966–969 (2005)
44. A Yannicopoulou, Visual aspects of written texts: preschoolers view comics. L1-Educ. Stud. Lang. Lit. **4**(2), 169–181 (2004)
45. M Tatham, K Morton, Expression in Speech: Analysis and Synthesis. Oxford Linguistics, (Oxford University Press, Oxford, 2006)
46. N Campbell, W Hamza, H Hoge, J Tao, G Bailly, Editorial special section on expressive speech synthesis. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1097–1098 (2006)
47. M Schröder, *Emotional Speech Synthesis—a Review, in Proceedings of EUROSPEECH*, vol. 1, 2001, pp. 561–5644
48. M Schröder, Expressive speech synthesis: past, present and possible futures, in *Affective Information Processing*, ed. by J Tao, T Tan, 2009, pp. 111–126
49. E Eide, A Aaron, R Bakis, W Hamza, M Picheny, J Pitrelli, A corpus-based approach to expressive speech synthesis, in *Proceedings of the 5th ISCA Speech Synthesis Workshop*, 2004, pp. 79–84
50. C Drioli, G Tisato, P Cosi, F Tesser, Emotions and voice quality: experiments with sinusoidal modelling, in *Proceedings of VOQUAL'03*, 2003, pp. 127–132
51. JF Pitrelli, R Bakis, EM Eide, R Fernandez, W Hamza, MA Picheny, The IBM expressive text-to-speech synthesis system for American English. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1099–1108 (2006)

52. M Theune, K Meijs, D Heylen, R Ordelman, Generating expressive speech for storytelling applications. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1137–1144 (2006)
53. WL Johnson, S Narayanan, R Whitney, R Bulut, M Das, C LaBore, Limited domain synthesis of expressive military speech for animated characters, in *Proc. IEEE Workshop on Speech Synthesis*, 2002, pp. 163–166
54. A Iida, N Campbell, F Higuchi, M Yasumura, A Corpus-Based, Speech synthesis system with emotion. Speech Commun. **40**(1), 161–187 (2001)
55. S Krstulovic, A Hunecke, M Schröder, An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements, in *Proceedings of InterSpeech*, 2007
56. M Schröder, J Trouvain, The German text-to-speech synthesis system MARY: a tool for research, development and teaching. Int. J. Speech Technol. **6**, 365–377 (2003)
57. A Abelin, J Allwood, Cross linguistic interpretation of expressions of emotions, in *Proceedings of the 8th Simposio Internactional de Communicacion Social*, 2003, pp. 387–393
58. KR Scherer, R Banse, HG Wallbott, Emotion inferences from vocal expression correlate across languages and cultures. J. Cross Cult. Psychol. **32**(1), 76–92 (2001)
59. M Pell, S Paulmann, C Dara, A Alasseri, S Kotz, Factors in the recognition of vocally expressed emotions: a comparison of four languages. J. Phon. **37**(4), 417–435 (2009)
60. F Burkhardt, N Audibert, L Malatesta, O Türk, L Arslan, V Auberge, Emotional prosody—does culture make a difference? in *Proceedings of speech prosody, Dresden, Germany*, 2006
61. P Boersma, D Weenink, Praat: doing phonetics by computer [Computer program] (2015). Version 5.4.22, retrieved 8 March 2016 from http://www.praat.org/
62. PJ Lang, M Bradley, B Culthbert, *International Affective Picture System (IAPS): Instruction Manual and Affective Ratings. Tech. Rep. A-6, the Center for Research in Psychophysiology* (University of Florida, Gainsville, Fla, USA, 2005)
63. D Tsonos, K Ikospentaki, G Kouroupetrglou, Towards modeling of readers' emotional state response for the automated annotation of documents, in *Proceedings of IEEE World Congress on Computational Intelligence, Hong Kong*, 2008, pp. 3253–3260
64. SG Nootboom, The prosody of speech: melody and rhythm, in *The Handbook of Phonetic Sciences*, ed. by WJ Hardcastle, J Laver, 1997, pp. 640–673
65. AC Rietveld, C Gussenhoven, On the relation between pitch excursion size and prominence. J. Phon. **13**(3), 299–308 (1985)
66. D Klatt, Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. J. Acoust. Soc. Am. **53**(1), 8–16 (1973)
67. D Pape, Is pitch perception and discrimination of vowels language-dependent and influenced by the vowels spectral properties? in *Proceedings of the International Conference on Auditory Displays*, 2005
68. J Bird, CJ Darwin, Effects of a difference in fundamental frequency in separating two sentences, in *Psychophysical and Physiological Advances in Hearing*, ed. by AR Palmer, A Rees, AQ Summerfield, R Meddis R (Whurr Publishers, London, 1998), p. 263
69. C Jones, L Berry, C Stevens, Synthesized speech intelligibility and persuasion: speech rate and non-native listeners. Comput. Speech Lang. **21**(4), 641–651 (2007)
70. J Romportl, J Kala, Prosody modelling in Czech text-to-speech synthesis, in *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, 2007, pp. 200–205
71. M Steedman, Information structure and the syntax-phonology interface. Linguist. Inq. **31**, 649–689 (2000)
72. X Huang, A Acero, HW Hon, Spoken Language Processing, a Guide to Theory, Algorithm and System Development. A guide to Theory, Algorithm and System Development (Prentice Hall, New Jersey, 2001)
73. D Braga, L Coelho, GR Fernando, *Subjective and Objective Assessment of TTS Voice Font Quality, in Proceedings of the 12th International Conference Speech and Computer*, 2007