Voice Driven Type Design

Matthias Wölfel
School of Digital Media
Furtwangen University
Furtwangen, Germany
Matthias.Woelfel@hs-furtwangen.de

Tim Schlippe
Research Karlsruhe
Karlsruhe, Germany
speech@research-karlsruhe.de

Angelo Stitz
School of Design
Pforzheim University
Pforzheim, Germany
info@metatype.de

***Abstract*—With *voice driven type design* (VDTD), we introduce a novel concept to present written information in the digital age. While the shape of a single typographical character has been treated as an unchangeable property until today, we present an innovative method to adjust the shape of each single character according to particular acoustic features in the spoken reference. Thereby, we allow to keep some individuality and to gain additional value in written text which offers different applications – providing meta-information in subtitles and chats, supporting deaf and hearing impaired people, illustrating intonation and accentuation in books for language learners, giving hints how to sing – up to artistic expression. By conducting a user study we have demonstrated that – using our proposed approach – loudness, pitch and speed can be represented visually by changing the shape of each character. By complementing homogeneous type design with these parameters, the original intention and characteristics of the speaker (personal expression and intonation) are better supported.**

Keywords—type design, typography, responsive type, speech analyzis, speech representation, adaptive character shape.

I. INTRODUCTION

The beginning of cultural evolution started with personal communication, first in oral and later also in written form. Both forms, oral and handwritten, do not only include the transfer of pure information, but are also a form of personal expression.[1] This, however, changed with the invention of using movable components (usually individual letters and punctuations) to reproduce the elements of a document. The world's first known movable type system was created in China around 1040 by Bi Sheng [2]. But not before the introduction of the movable-type printing system in Europe by Johannes Gutenberg around the 1450s [3], movable types could demonstrate their superiority. In contrast to the thousands of characters needed in the Chinese writing system, European languages need a much lower number which makes it much easier to handle. After their invention in the 1860s, typewriters became a convenient tool for practically all written communication and quickly replaced handwriting except for personal correspondence [4]. While industrialization necessitated standardization of type in that replication process and their materials [4], digitization offers a liberation of these stringent formats: Fonts have been developed in all kind of flavors. But keys, since the invention of the typewriter, stayed the preliminary input modality. This has also not been changed in the late 1960s by the invention of the word processor [4]. Due to this restriction of keys which provide only two states[2] on/off and time information, we are not able to express individual characteristics.

These individual characteristics, intonation as well as the emotional state can then only be 'reconstructed' either by explicit reference (e.g. "I'm happy!") or other verbal (e.g. linguistic behavior such as changes in disagreement, affect terms, and verbosity) or nonverbal cues (e.g. use of punctuation or emoticons [5]) [6]. Speech as an alternative form of input modality, in contrast to a keyboard, contains more information. This is complementary to text and reflects individuality and emotion. However, it is simply thrown away and not used to influence the type design or represented in other graphical form. We got so used to this generic transfer of information that nobody challenges this way of representing information. What gets lost becomes more obvious, for instance, if we transform written text back to speech with 'simple' speech synthesis[3]. Without emotion, prosody and personalization listening to speech gets very boring very soon.

In order to keep the additional information present in verbal communication also in text-based communication, we introduce a novel concept: In *voice driven type design* (VDTD) the shape of a single character adjusts according to particular acoustic features in the spoken reference, such as loudness, pitch, and speed.

II. RELATED WORK

Using typography as a stylistic device has a very long tradition which comes in different flavors. It can be:

- *static* such as presented in printing books, posters and comics, where type represents content in a uniform and permanent way [7],

[1] See for example the controversially discussed field of graphology which is interested in analyzing the physical characteristics and patterns of handwriting to identify the writer, indicate the psychological state at the time of writing, or evaluate personality characteristics [1].

[2] In contrast to keyboards for text input, musical keyboards offer a range of expression types including velocity sensitivity (how fast a key is pressed), pressure sensitivity (amount of force on a held-down key) and displacement sensitivity (distance that a key is pressed down).

[3] With 'simple' we refer to speech synthesis which does not use linguistic analysis to estimate intonation and duration.

- *dynamic* as in kinetic typography which is an animation technique to express ideas using text-based video animation [8, 9], or
- *reactive* as in responsive type [10] where the shape of each letter is adjusted according to the properties (such as age, eye sight, relative position, and speed) of the reader.

While the previously mentioned approaches, in general, use a uniform character style, *sound poetry* ignores these constrains. Sound poetry is probably best described as an artistic form of "performing" a written text focusing on the phonetic aspects of speech besides its semantic values [11]. For instance, each line of the poem *Karawane*, by one of the pioneers of Dadaist poetry Hugo Ball [12], reflects – through different typographic styles – "the acoustic dimension of a linguistic sign" [13].

Another approach, instead of manipulating the look of the text itself, is to use *emoticons* – yet to express emotions instead of an acoustic dimension. Even though emoticons have been demonstrated to be effective for remote emotional communication [14], additional characters have to be added and variation within text cannot be represented well. Another drawback of this approach is that not all emoticons are interpreted equally between different cultures [15].

Automatic processing of speech is a research topic with a rather long tradition. However, speech as well as emotion recognition on acoustic features have been and still are treated as independent entities. On one hand, automatic speech recognition systems are still limited to recognizing what has been said without being concerned about how. On the other hand, emotion recognition systems aim to classify the type of emotion that lies within the acoustic speech signal. How those emotions could be expressed in text-based communication has not been widely investigated [16].

Using acoustic cues to determine the look of text or additional hints such as emoticons, to the best of our knowledge, has only been started to be investigated recently: Zimmermann [17] proposed to select pre-existing emoticons from multimedia data (including video, still image, and/or audio) captured by a device. Furthermore, he also proposed to generate emoticons based on the expressions on the user's face. Matsumiya et al. [16] have proposed and investigated how to automatically generate the shape and appearance of text balloons based on linguistic and acoustic speech features. Given a chance rate of 73% to estimate the shape of text balloons on their comics-anime test corpus, they reached 87% accuracy and demonstrated that subtitling with text balloons is better than that with static text.

III. Voice Driven Type Design

As described in the previous section, related work has been limited to a sentence-based selection of appropriate text fragments corresponding to spoken utterances. In contrast, the focus of our work is to present and investigate a granularity with a much higher resolution, namely on each single phoneme and their corresponding grapheme or graphemes, respectively: We propose to vary the shape of a single character according to particular acoustic features in the spoken reference. Our motivation of a grapheme-level adaptation of the transcription is to better represent the characteristics of the spoken utterance and to keep individuality in written text. This strategy allows additional meta-information in subtitles and chats, supporting hearing impaired and deaf people, illustrating intonation and accentuation in books for language learners and giving hints how to sing. These would be more limited or even not possible at all with previous approaches.

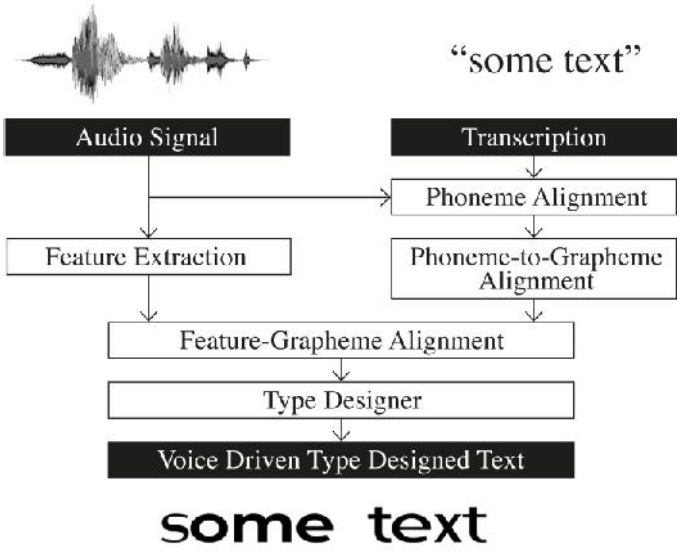

Fig. 1. From speech signal to voice driven type designed text.

A. Algorithm

As illustrated in Figure 1, to retrieve VDTD given a spoken utterance and its transcription[4] consists of the following steps:

1) Phoneme alignment and acoustic feature extraction
- Generate phoneme transcription and determine the beginning and end of each phoneme.
- Determine *loudness* and *pitch* (step size 10 ms) given the acoustic signal.

2) Phoneme-to-grapheme alignment
- Align each phoneme or phoneme sequence to one or more graphemes.

3) Features-grapheme alignment
- Determine the *speed* parameter of each grapheme by using the beginning and end times of the corresponding phoneme or phoneme sequence.
- Determine the *loudness* and *pitch* parameters of the grapheme by averaging *loudness* and *pitch* according to the beginning and end times of the corresponding phoneme or phoneme sequence.

4) Type design

[4] The transcription can either be given a-priori or automatically generated with automatic speech recognition.

- Generate the shape of each character according to the corresponding normalized (mean and variance) and mapped features *loudness*, *pitch* and *speed*.

To retrieve the phoneme transcription and to align the audio sequence to the word sequence, we used the Munich Automatic Segmentation System MAUS [18]. The acoustic features were extracted using our own code which analyzes weighted Fourier spectra to get loudness and cross-correlation to determine pitch. To align the phoneme sequence to the grapheme sequence, we applied the m2m-aligner [19].

1) Normalizing the speech parameters

The mean and variance values of the acoustic features vary between the phonemes and phoneme classes. For example, a vowel is significantly louder than a fricative and has a wider range in loudness. Since the goal of our proposed approach is to visualize acoustic variation and not to project such differing ranges across phonemes and phoneme classes, we need to compensate for the different means and variances per phoneme before applying the features. Furthermore, the generated typographical character sequence would consist of uneven distributed characteristics per phoneme class. This would, for instance, result in a sequence where all vowels would always be displayed with a wider stroke than fricatives.

By normalizing each phoneme class (c) according to

$$\mathbf{p}_c = 0.5 + 0.25 * (\mathbf{p}_c - \boldsymbol{\mu}_c)/\boldsymbol{\sigma}_c;$$

where ($\mathbf{p}$) denotes the acoustic parameters, ($\boldsymbol{\mu}$) represents the mean values and ($\boldsymbol{\sigma}$) the standard deviation, a homogeneous design where only the variation of each phone per class is pronounced can be provided. The mean and standard deviation are calculated from a training set consisting of various utterances by different speakers. To guarantee that the resulting parameters lie in the range of 0 and 1, the values are normalized. Then the normalized acoustic parameters (*loudness*, *pitch* and *speed*) of a phoneme are handed over to the corresponding grapheme or grapheme sequence to form the graphical parameters (*vertical stroke weight*, *horizontal stroke weight,* and *character width*).

2) Correspondence between phonemes and graphemes

The relationship between graphemes and phonemes varies among languages [20, 21]. Languages with alphabetic scripts are characterized by four types of relationships between phonemes and graphemes: a 1-to-1, a *p*-to-1, a 1-to-*g* and a *p*-to-*g* mapping, where *p* and *g* are integer values greater than 1. Therefore, in the simplest case, one phoneme represents one grapheme. In all other cases a 1-to-1 relationship is not given.

In case of a close grapheme-to-phoneme relationship, such as German, mostly one character represents one phoneme. Consequently, each character can be adapted based on the acoustic characteristics of the corresponding phoneme. To deal with all kind of phoneme-to-grapheme relationships, we apply the following strategy which is also demonstrated in Figure 2:

1) Perform an automatic forced alignment process which aligns one phoneme to one up to three corresponding graphemes [22].
2) Transfer the characteristics of the phoneme to the corresponding grapheme.

For German and English, a mapping of a phoneme to up to three characters has proved to be successful (e.g. for English igh - /aɪ/ and sh - /ʃ/). Diphthongs (e.g. /aɪ/, /ɪə̯/ and /aʊ/) are handled as one phoneme. However, this mapping can be easily adapted according to the grade of relationship of new target languages.

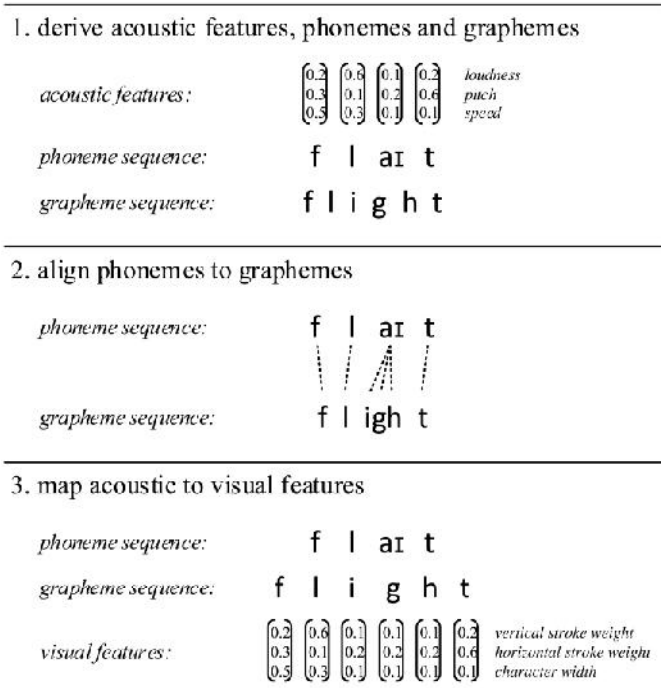

Fig. 2. Mapping the acoustic features of a phoneme sequence to the visual features of the corresponding grapheme sequence.

B. Visualization

Today's typefaces hold only a limited range of mostly nine fonts depending on different nuances for stroke weight (light, regular, black) and character width (extended, regular, condensed). However, if we want to map continuous characteristics of the voice into a visual representation, we do not only need mapping functions for the different acoustic parameters, but a continuous visual representation. Therefore, an apparatus to change the character on a continuous scale is required. In addition, we have the requirement that the proposed approach has to work for longer text passages which constrains the degree of freedom in our visual expression. For now we have decided to work with the parameters *vertical stroke weight*, *horizontal stroke weight* and *character width*. The three free parameters are demonstrated in Figure 3.

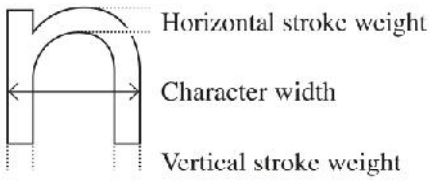

Fig. 3. Demonstrating the three freely adjustable parameters – *vertical stroke weight*, *horizontal stroke weight* and *character width*.

Other possible parameters such as *height*, *contrast*, *sharpness*, and *skewness* have been decided to not be considered for the reasons just given. But other freely adjustable parameters, of course, can be considered in future work.

To express a bandwidth of emotional states through written text leads to the conclusion that the origin point of a dynamic character shape must constitute a simple, generic, rational and reduced form. Therefore, we adopted one of the most satisfactory, modernist sans serif typefaces of the twentieth century "Futura" by Paul Renner (1927) [23, p. 80].

1) Continuous scale of character shape

Since, to the best of our knowledge, no type family or software exists which is able to fulfill our particular needs, we designed our own type family and developed a font processing tool. This enables to change every parameter of each character in real time without loosing distinctive and aesthetic character shapes. To guarantee a functional, stringent and aesthetic character shape with our automatic mapping function, we manually defined the extrema of our continuous space as demonstrated in Figure 4 with the three type design dimensions – *vertical stroke weight*, *horizontal stroke weight* and *character width*.

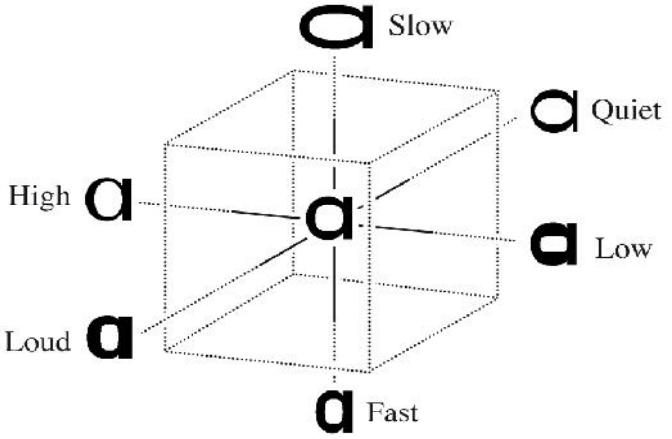

Fig. 4. Continuous interpolation between each parameter. Each parameter runs in a range of 0 and 1. The values of the parameters are represented in a triplet (*vertical stroke weight*, *horizontal stroke weight, character width*). The average font (0.5, 0.5, 0.5) is located in the center of the space.

2) Mapping voice characteristics to character shape

Every single piece of information sent and perceived by humans is embedded in a particular manner of formatting which goes far beyond the transfer of pure information. Adding values present in acoustic signals into a visual representation makes only sense if it can be interpreted to 'extract' the original meaning. Thus, a comprehensible and reasonable relationship between speech characteristics and the character shape has to be found. This requires that common principles in formatting exist in speech and typography and that these principles can be well mapped. After these considerations, we decided to map the voice to the character shape as follows:

- *Loudness:* Producing loudness in speech amplifies the signal and is usually used to have the attention of a listener. To have the attention of the reader, bolder text – produced with more stroke weight – is commonly used since it makes it easier and more efficient to scan the text and recognize important keywords [24]. Increasing the stroke weight commonly effects the vertical and horizontal stroke weight equal. To recognize each acoustic feature separately after the mapping on its visual representation, we decided to increase only the vertical stroke weight. Contrary to the adjustment of the horizontal stroke weight, increasing the vertical stroke weight is more common and attracts the attention of the reader which can be explained with the historical development of the classified styles of types [25].
- *Pitch:* Numerous studies confirm that emotional expression of utterances is formed by variations of pitch levels [26, 27, 28]. High pitch levels draw attention of a listener and express additional emotions, such as joy, anxiety or fear, while medium pitch levels account for more neutral attitudes [29]. While we adjust the vertical stroke weight according to the loudness, we adapt the horizontal stroke weight depending on the pitch level since this modification increases the reader's curiosity. Due to this aspect, the inverse-contrast (Italian) [30] fonts with significant vertical stroke weight have a high recognition factor. Nicolete Grey wrote about these Italian typefaces: "a crude expression of the idea of perversity" [31], while others call it "degenerated" [32]. Consequently, we learn that adjusting the horizontal stroke weight touches the reader's emotions and fits to express pitch.
- *Speed:* The processes of information transfer with speech and reading happen within a time period. A reader usually jumps from a part of a word to a next part of a word [33]. Increasing the character width extends this scanning process of the eyes. Therefore, we map the speed of the utterance to the character width.

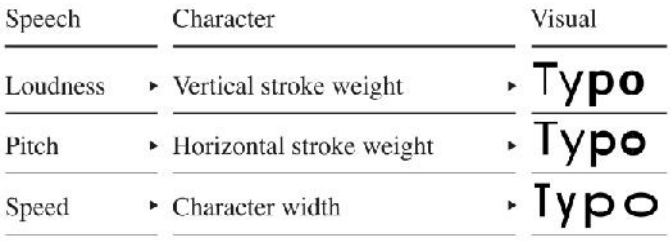

Fig. 5. Mapping speech characteristics on text formatting.

Figure 5 summarizes the mapping of the acoustic characteristics *loudness*, *pitch* and *speed* to its visual representations *vertical stroke weight*, *horizontal stroke weight* and *character width*.

IV. APPLICATIONS

This section presents some possible applications where VDTD can be used to support given text by providing additional information.

A. *Language learning and speech-language pathology*

Many people learning a new language have trouble doing the intonation and accentuation in a right way. A potential of VDTD is that it allows to illustrate intonation and accentuation in software and books for language learners.

Fig. 6. Language learning with VDTD.

Figure 6 illustrates the VDTD representation of the Spanish word "attencíon" meaning "attention". While the "accented i" indicates an intonation of the "i" in the static text, one has no glue about the intonation in the beginning of the word. However, with the help of VDTD, the stress of the "a" is conspicious.

In addition to language learning, VDTD indicates in other fields how to pronounce words, e.g. in speech pathology. It enables to catch additional information which may not be possible with exclusively acoustic information. For instance the change of loundness within a spoken word migth not be recognizided by the listener, but if he sees its VDTD transcription he might be able to see the differences.

B. *Hints for deaf people*

Deaf people profit from VDTD since it gives them hints on how something was spoken. This is helpful in different situations: (1) learning how to pronounce (similar to language learning and speech-language pathology) or (2) interpreting how something is intended.

There have been several efforts – even by big organizations such as BBC and Amnesty International – to make the life after hearing loss more comfortable (e.g. The Future of Subtitling[5]). Our approach has the potential to enrich the television and cinema experience for deaf people.

C. *Hints for dyslexia*

A reading disorder is primarily influenced by the so-called *phonological awareness* [3]. It refers to an individual awareness to the phonological structure of words [35]. Phonological awareness involves the detection and manipulation of sounds at three levels of sound structure: (1) syllables, (2) onsets, rimes and (3) phonemes [36] and has also an impact during the process of reading and writing [37]. People with an unsatisfactory phonological awareness are not able to extract the correct orthographic word out of a spoken utterance. VDTD offers a new way to visualize a comprehensible relationship between the spoken utterance and the written text.

D. *Subtitles*

Subtitles on television screens are very popular in places with either a lot of background noise (in the gym, on stations, at the airport) or where different programs are simultaneously broadcasted (in sports bars where several football games, baseball games or basketball games are shown). They are switched on in the gym, in sports bars, on stations, at the airport and other places. Figure 6 shows a scene during a touchdown – a very emotional experience at least for someone who is excited in this sport. While a static subtitle does not transfer the TV host's emotion, the subtitle modified with our approach reflects the host's admiration to carry the ball for over 80 yards and his enthusiasm.

Fig. 7. Scene of a football game with VDTD subtitles.

E. *Texting*

People love texting – be it on the smartphone, on Twitter, during gaming, etc. Additionally, using automatic speech recognition to automatically generate the text message and sending voice messages has become more and more popular. Emoticons help us to add meta-information transferring irony and emotions. If it is not possible to listen to the voice messages, there are already services transcribing them and sending them back in text format. However, in the static text of the transcription, meta-infromation is lost and no emoticons help the receiver. VDTD can support such a scenario.

Fig. 8. Chatting with the support of VDTD text.

Figure 8 demonstrates a snippet of a chat. Given only static text it is not clear if Rajat is serious with his question about their Word Cup's plans. Using VDTD could help here to decode irony as meta-information into the question.

[5] http://www.actiononhearingloss.org.uk

F. Singing and karaoke

Voice-driven type designed songbooks can support singing songs without the requirement of being able to read notes since *loudness*, *pitch* and *speed* are illustrated in a more comprehensible form and provides additional information to those who are able to read and interprete notes.

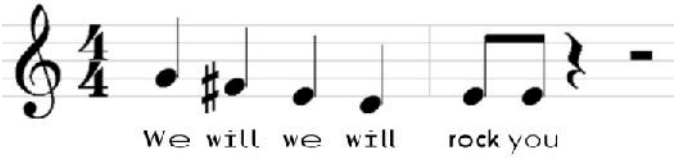

Fig. 9. Singing with VDTD (Song: Queen – We will rock you).

Figure 9 and 10 demonstrate the refrain of Queen's song "We will rock you". Using VDTD the pitch decrease of the first four words is visible. Additionally, we have information about the loudness which is mostly not represented in notes like in this example. With the proposed approach it can be observed that the word "rock" has to be sung much louder. VDTD also provides information with a more detailed resolution: While the notes provide a single tone a whole word is sung. Given only the text it is unclear how long each phoneme has to be pronounced. In the given example we see that the 'e' in both 'we's and the 'ou' in 'you' have to be pronounced longer than the other phonemes.

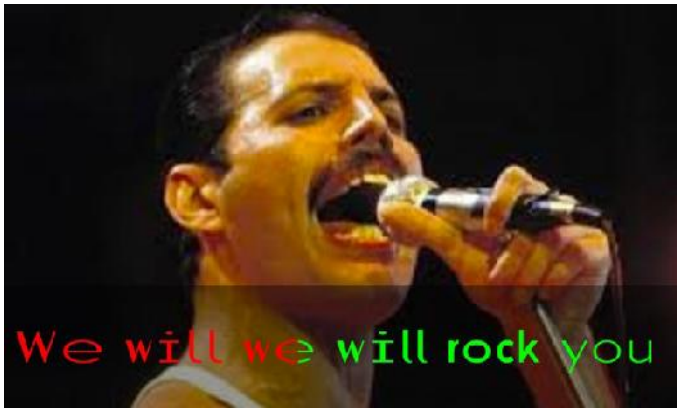

Fig. 10. Karaoke TV screen using VDTD (Song: Queen – We will rock you).

VDTD is also helpful to indicate how to sing in Karaoke. This is demonstrated in Figure 10 where the change beween red text to green text indicates the current position of the singer.

V. Evaluation

To evaluate our approach, we performed a user study to answer

- whether people agree that VDTD can be used for several applications.
- whether *loudness*, *pitch* and *speed* of the original speech signal are represented well by the mapping function.
- whether the proposed visual representation can represent emotions.
- whether speaker-specific characteristics can be observed.

A. Experimental Setup

For our experiments, we asked four persons to read Wolfgang Goethe's German poem "Erlkönig" (Erl-King) [38][6]. We used close-talking microphones to provide recordings without distortions which could reduce the reliability of the extracted acoustic features [39]. In order to investigate similarity patterns for individual speakers, we recorded each speaker twice. Therefore, our dataset consists of 8 audio files from 4 speakers.

The participants of our user study were randomly selected volunteers who participated free of charge. In total, 9 participants (6 female, 3 male) were exposed to VDTD and asked to fill out a questionnaire. 5 subjects were in the age range between 20 and 30 years, 1 subject between 30 and 40, 1 between 40 and 50 and 2 above 50. 6 subjects were design professionals or students, and 3 had no particular design background. Our goal was to have both, participants with more sense of typography and participants with less sense of this topic. The participants' mother tongue is German. They understand well the tragic poem with its emotional aspects, most of them know it from school. Except for the survey which will be described in Section E, the score range, followed the rules of a forced choice Likert scale, which ranged from (1) *strongly agree*, (2) *agree*, (3) *neutral*, (4) *disagree* to (5) *strongly disagree*.

B. Acceptance

The question if VDTD is interesting for several applications has been answered by most participants positively with an average score of 1.5 – a result between strongly agree and agree. They see its applications in particular in poems, in learning a second language, in providing hints for acting and in artistic expression.

C. Representation of Voice Characteristics

In this experiment, subjects listened to one recording of the poem and graded how good each verse of five texts represents the *loudness*, *pitch* and *speed* of the recording. Table 1 gives the results – VDTD: our algorithm has been used; R1 and R2: randomly chosen parameters; and HTD: homogenous type design (static text).

We see that our approach results in the best scores and that *loudness* and *pitch* are better represented than *speed*. Random variation result also in positive evaluations in contrast to HTD. This can be explained by *clustering-illusion* [40] and *Apophenia* [41] – humans seek sense in random patterns.

In this context it is also interesting to note that those volunteers who have a strong linguistic background compared

[6] "Erlkönig", written in 1782, is a very emotional poem with a tragical ending consisting of 8 verses with each 4 lines. The length of a recording of the 32 lines is approximately 2 minutes on average.

and analyzed the different visualizations more carefully. Some discovered that, in the case of random changes, the visualization is not consistent to the way people pronounce the given words.

Acoustic Feature	*Approach*			
	VDTD	***R1***	***R2***	***HTD***
Loudness	2.46	2.71	2.81	3.00
Pitch	2.46	2.58	2.40	3.10
Speed	2.58	2.56	2.83	3.23
Average	2.50	2.62	2.68	3.11

TABLE I. SCORES OF THE TEXTUAL REPRESENTATION OF THE SPEECH CHARACTERISTICS.

D. Representation of Emotions

In this experiment, the subjects were asked how each text reflects the emotions expressed. The subjects were given three texts as shown in Figure 11 – 1: represents homogenous type design and got an average score of 4.44; 2: represents the random approach and got an average score of 2.67; and 3: represents VDTD and got an average score of 2.33. It is interesting to note that people slightly agree that emotions are present even though the parameters are randomly chosen. Our approach, however, is agreed to represent more information.

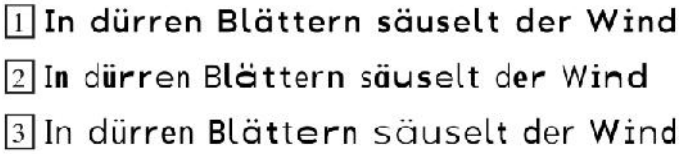

Fig. 11. Example of text generated by homogenous, random and voice driven parameters.

E. Representation of Speakers' Characteristics

To determine if the visualization representation preserves individual characteristics of a speaker, we have calculated the standard deviation of the visual variation, represented by the acoustic features, between utterance pairs of the same speaker as well as from two different speakers. Comparing the standard deviations in Table 2 of the three normalized features (to range between 0 and 1) *loudness*, *pitch* and *speed*, we see that all of them are significantly smaller for the same speaker in contrast to different speakers.

Acoustic Feature	*Standard Deviation*	
	Same Speaker	***Different Speaker***
Loudness	0.028	0.040
Pitch	0.032	0.045
Speed	0.019	0.027

TABLE II. STANDARD DEVIATION OF THE TEXTUAL REPRESENTATION OF ACOUSTIC FEATURES FOR THE SAME SPEAKER (COMPARING TWO RECORDINGS OF THE SAME UTTERANCE) AS WELL AS FOR OTHER SPEAKERS.

But are these parameters represented well in the shape of the characters? We asked our participants to find four utterance pairs given 8 utterances from 4 speakers, see Figure 12. The results in Table 3 show that four participants found all four pairs, a probability of 1 to 105 each. We, therefore, can conclude that the visual representation gives some clue about speaker specific properties.

Fig. 12. Individual speech characteristics are also visible in spoken text due to VDTD. Utterance pairs of the same speakers (1,2), (3,4), (5,6), and (7,8).

Number of Utterance Pairs	4	2	1	0
Found x Times	4	3	1	1

TABLE III. FINDING PAIRS WITH THE HELP OF SPEECH CHARACTERISTICS VISIBLE IN TEXT – DISTRIBUTION OF FOUND UTTERANCE PAIRS

VI. CONCLUSION AND FUTURE WORK

We have proposed VDTD as a novel concept to represent additional information present in verbal communication also in text-based communication. To keep and convert this information, we change the shape of each single character according to particular acoustic features.

Potential applications of our proposed approach to visualize the characteristics present in the voice in the type of the

characters are manifold: It has the potential to support learning to read and speaking a foreign language. It can provide hints for actors on the intended prosody. Furthermore, it offers novel possibilities in subtitles on television screens – which might be particularly beneficial for hearing impaired and deaf people. VDTD can be helpful in singing and karaoke and adds meta-information to transcribed voice messages.

In addition to the experiments described in this paper, we have gained positive experiences and feedback with our interactive VDTD demo system which we presented at the exhibition *GLOBALE: Infosphere*[7] hosted by ZKM Karlsruhe (Center for Art and Media) [42] and at the conference *Mensch & Computer (MuC 2015)* [43] in Germany. In this system we let the audience read poems aloud and then automatically provide them the voice driven type designed text on a screen.

Having presented and discussed potential VDTD application scenarios and completed initial analyses of acceptance and information gain, we plan further analyses and usability studies to find optimal features and conditions for the applications. For example, we have demonstrated that the modification of vertical and horizontal stroke weight plus character width works to give the reader hints about the *loudness, pitch,* and *speed* of spoken utterances. How strong the shape of the character has to be changed has not been investigated so far and might be even depending on the application. For instance, for characters in subtitles of serious anker speakers, a smaller interpolation space of the characters may be chosen than for the characters of subtitles in football games, baseball games or basketball games to represent more emotions. Besides the context, the strength of the changes may also depend on the font size. Another interesting question is if the change in character shapes has to be evident to the speaker or if it can already support reading if the changes are so small that they are not consciously recognized.

Future work may include a comparison and analysis of the acceptance of other modifications in the typography depending on the application scenario, e.g. the sharpness to represent pitch variations. We have transferred the characteristics in the voice at the character level since our intention was to be able to represent differing spoken characteristics in different positions of the word. For certain applications or languages it may be better to turn to a syllable or word level.

We have *evaluated* our experiments with German, our mother tongue, – a Germanic language with a fairly regular grapheme-to-phoneme relationship [20]. Future work will contain additional evaluations to languages with a more ambiguous relationship, e.g. English, where one letter can represent a variety of sounds conditioned by complex rules and many exceptions [44]. Further work could also include the application to other writing systems such as Arabic, Hebrew, Greek, Kyril, Hangul, Hiragana and Katagana. How to adapt VDTD to logograms and pictograms as used in Chinese characters could be another interesting question to be approached.

A further challenge is to tackle numbers. To generate the corresponding phoneme sequence, a sequence of digits needs to be transferred to characters with the help of number spellers. Then, we can adapt the characters with our approach. However, how to transfer the voice characteristics to the single digits? This is a challenge since much more phonemes correspond to single digits than in the case with words composed of characters (e.g. 24 = twenty four). Moreover, in contrast to words, numbers are not pronounced from left to right consistently in all languages. For example, in German the number "24" is pronounced as "four-and-twenty".

Further topics are the adaptation of punctuation marks and whitespaces between word tokens. Is it beneficial to modify the shape of a comma, question mark, exclamation mark, etc. according to the characteristics of the previous character or not at all? For aesthetics it can make sense to transfer the characteristics of the previous character. So far we have not adapted the whitespace to the duration of the break between two words. However, an adaptation according to the time information is promising, for example to display breaks in karaoke applications.

So far we have applied our method on sans serif letters. Another possible application would be to apply it to imitate handwriting or calligraphy. Using VDTD for the automatic production of handwriting and calligraphy may lead to a more realistic type face variation in contrast to random or rule-based variations of the type face.

VII. REFERENCES

[1] J. Berger, Handwriting Analysis and Graphology. In M. Shermer (ed.) The Skeptic Encyclopedia of Pseudoscience, ABC-CLIO, pp. 116-120, 2002.

[2] J. Needham and T. Tsuen-Hsuin, Science and Civilisation in China: Volume 5, Chemistry and Chemical Technology, Part 1, Paper and Printing, ser. Science and Civilisation in China. Cambridge University Press, 1985.

[3] R. Kinross, Modern Typography, vol. 2, London: Hyphen Press, 2010.

[4] E. C. Berkeley, "Secretaries Get a Computer of their Own to Automate Typing, Computers and Automation," in Computers and Automation, vol. 18, no. 1, pp. 59, 1969.

[5] J. B. Walther and K. P. D'Addario, "The impacts of emoticons on message interpretation in computer-mediated communication," Social science computer review, vol. 19, no. 3, pp. 324-347, 2001.

[6] J. T. Hancock, C. Landrigan, and C. Silver, "Expressing emotion in text-based communication," in Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp. 929-932, 2007.

[7] P. Shaw, "Codex: The Journal of Letterforms," in The Menhart Issue. John Boardley, 2012.

[8] J. Lee, S. Jun, J. Forlizzi, and S. E. Hudson, "Using Kinetic Typography To Convey Emotion In Text-Based Interpersonal Communication," in The 6th conference on Designing Interactive systems. ACM, pp. 41-49, 2006.

[9] R. Rashid, Q. Vy, R. Hunt, and D. I. Fels, "Dancing with Words: Using Animated Text for Captioning," Intl. Journal of Human – Computer Interaction, vol. 24, no. 5, pp. 505-519, 2008.

[10] M. Wölfel and A. Stitz, "Responsive Type—Exploring Possibilities of Self-Adjusting Graphic Characters," in Proceedings of Cyberworlds 2015, 2015.

[11] M. Perloff and C. Dworkin, The sound poetry / the poetry of sound, University of Chicago Press, 2009.

[12] R. Huelsenbeck, Dada Almanach, Berlin, Erich Reiss Verlag, pp. 53, 1920.

[13] E. Adamowicz and E. Robertson, DaDa and Beyond, vol. 1, DaDa Discourses, Editions Rodopi B.V., Amsterdam – New York, pp. 42, 2011.

[7] http://zkm.de/event/2015/09/globale-infosphare

[14] L. L. Rezabek and J. J. Cochenour, “Visual Cues in Computer- Mediated Communication: Supplementing Text with Emoticons,” vol. 18, Journal of Visual Literacy, no. 2, 1998.

[15] B. Mesquita, and R. Walker, Cultural differences in emotions: A context for interpreting emotional experiences, Behaviour Research and Therapy, 41(7), pp. 777-793, 2003.

[16] S. Matsumiya, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, “Data-Driven Generation of Text Balloons based on Linguistic and Acoustic Features of a Comics-Anime Corpus,” in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[17] R. Zimmermann, “Use Of Multimedia Data For Emoticons In Instant Messaging,” Jul. 28 2005, US Patent App. 10/767,132. [Online]. Available: https://www.google.com/patents/US20050163379

[18] T. Kisler, F. Schiel, and H. Sloetjes, “Signal Processing Via Web Services: The Use CaseWebMAUS,” in Digital Humanities 2012, Hamburg, Germany, pp. 30-34, 2012.

[19] S. Jiampojamarn, G. Kondrak, and T. Sherif, “Applying Manyto- Many Alignments and Hidden Markov Models to Letter-to- Phoneme Conversion,” in Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, New York: Association for Computational Linguistics, pp. 372-379, April 2007.

[20] T. Schlippe, S. Ochs, and T. Schultz, “Grapheme-to-Phoneme Model Generation for Indo-European Languages,” in the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), Kyoto, Japan, 25-30 March 2012.

[21] M. Goudi and P. Nocera. “Sounds and Symbols: An Overview of Different Types of Methods Dealing With Letter-To-Sound Relationships In A Wide Range Of Languages In Automatic Speech Recognition,” in Proceedings of SLTU 2014, St. Petersburg, Russia, 14-16 May 2014.

[22] A. W. Black, K. Lenzo and V. Pagel, “Issues in Building General Letter to Sound Rules,” ESCA Workshop on Speech Synthesis, 1998.

[23] R. Poulin, Graphic Design and Architecture, A 20th Century History: A Guide to Type, Image, Symbol, and Visual Storytelling in the Modern World. Rockport Publishers, pp. 80, 2012.

[24] R. Bringhurst, The Elements of Typographic Style, vol. 3.2, Hartley and Marks Publishers, pp. 55-56, 2008.

[25] D. B. Updike, Printing types, their history, forms and use, a study in survivals, Geoffrey Cumberlege, vol. 2, Oxford University Press, London, 1922.

[26] M. Pell, M. Paulmann, S. Dara, A. Alasseri, and S. Kotzb, Factors in the recognition of vocally expressed emotions: a comparison of our languages, J Phon, vol. 37, pp. 417-435, 2009.

[27] O. M. Bachorowski J, Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context, vol. 6, Psychol Sci, pp. 219-224, 1995.

[28] C. Williams and K. Stevens, Emotions and speech: some acoustical correlates, J Acoust Soc Am, vol. 52, pp. 1238-1250, 1972.

[29] E. Rodero, “Intonation and emotion: Influence of pitch levels and contour type on creating emotions,” vol. 25, no. 1, Journal of Voice, pp. 25-34, 2011.

[30] Caslon & Catherwood’s, Type specimen, 1821.

[31] N. Gray, Nineteenth Century Ornamented Typefaces, Faber & Faber Ltd, London, 1938.

[32] J. H. Benson and Carey, Arthur Graham, The elements of lettering, Newport, Rhode Island: John Stevens, 1940.

[33] G. Unger, Wie man’s liest, Sulgen, Zürich: Niggli Verlag AG, pp. 63-65, 2006.

[34] S. Phillips, K. Kelly, L. Symes, Assessment of Learners with Dyslexic-Type Difficulties, SAGE, pp. 7, 2013.

[35] W. E. Tunmer & C.M. Fletcher, The Relationship between Conceptual Tempo, Phonological Awareness, and Word Recognition in Beginning Readers, Journal of Literacy Research, vol. 13, no. 2, pp. 173-185, 1981.

[36] J. B. Gleason, e-Study Guide for: The Development of Language, Content Technologies, 2012.

[37] C. Schnitzler, Phonologische Bewusstheit und Schriftspracherwerb, Georg Thieme Verlag, Stuttgart, pp. 1, 2008.

[38] “Der Erlkönig,” 2014, http://en.wikipedia.org/wiki/Der Erlkönig.

[39] M. Wölfel and J. McDonough, Distant Speech Recognition, John Wiley & Sons, 2009.

[40] R. D. Clarke, “An application of the poisson distribution,” vol. 72, Journal of the Institute of Actuaries, p. 481, 1946.

[41] K. Conrad, Die beginnende Schizophrenie. Versuch einer Gestaltanalyse des Wahns, Stuttgart: Georg Thieme Verlag, 1958.

[42] M. Wölfel, A. Stitz, and T. Schlippe, „Voice Driven Type Design,” GLOBALE: Infosphere, ZKM (Center for Art and Media), Karlsruhe, Germany, 2015-2016.

[43] M. Wölfel, A. Stitz, and T. Schlippe, „A Voice Driven Type Design Demo,” in Proceedings of Mensch und Computer 2015, pp. 413-416, Stuttgart, Germany, 2015.

[44] A. Waibel, H. Soltau, T. Schultz, T. Schaaf, and F. Metze, “Multilingual Speech Recognition,” Verbmobil: Foundations of Speech-to-Speech Translation, ed. Wolfgang Wahlster, Springer Verlag, 2000.