Discrimination and Perception of the Acoustic Rendition of Texts by Blind People

Vassilis Argyropoulos[1], Konstantinos Papadopoulos[2], Georgios Kouroupetroglou[3], Gerasimos Xydas[3], and Philippos Katsoulis[3]

[1] University of Thessaly, School of Humanities, Department of Special Education,
Argonafton & Filellinon St, Volos, GR-38221, Greece,
vassargi@uth.gr
[2] University of Macedonia, Department of Education and Social Policy,
kpapado@uom.gr
[3] University of Athens, Department of Informatics and Telecommunications,
Panepistimiopolis, GR-15784, Athens, Greece
{koupe,gxydas,}@di.uoa.gr

Abstract. This paper reports on the results from a series of psychoacoustic experiments in the field of the auditory representation of texts via synthetic speech which comprise similar acoustic patterns so called "paronyms". The errors which occur when listening to paronyms are classified as errors of phonological type. Thirty blind and thirty sighted students participated in psychoacoustic experiments. The results from the experiments depicted the types of the subjects' errors and addressed comparisons between the performances of blind and sighted students on their auditory distinctions towards the chosen scripts (paronym words and sentences with paronyms). The discussion considered the practical implications of the findings such as issues regarding education as well as the development of suitable design of acoustic rendition of texts in favor of better perception and comprehension.

Keywords: text-to-speech, synthetic speech perception, discrimination of synthetic speech, paronym words.

1 Introduction

Related works in the field of the acoustic representation of documents and texts (text-to-speech and document-to-speech) have pointed out the need for rendering with distinction peculiar linguistic patterns – such as paronym words - and part of the meta-information [1], [2], [3], [4].

More specifically, paronym words are characterized as linguistic structures which their acoustic representation bears a great resemblance between them and as such they might create misunderstandings or confusions to listeners. For example, the words "affect" → [af`ekt] and "effect" → [ef`ekt] in English and the words "φίλη" → [f`ili] and "φυλή" → [fil`i] in Greek are considered as paronyms. There are two reasons behind these misunderstandings when listening to paronyms. The first reason originates from the similarity of the phonemic representation of paronymal words. For

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2007, LNCS 4556, pp. 205–213, 2007.

example: “έκλειψη” → [`eklipsi] and “έκθλιψη” → [`ekTlipsi]. The acoustic realization of these words can be confusing for the listener, since they only differ in the unvoiced fricative [T] that might not be heard (when saying [`ekTlipsi] or might be thought as being heard (when saying [`eklipsi]). Speech rhythm is a significant factor that affects this perception.

The second reason of the misunderstandings originates from the intonational realization of the phrases. There are cases where though the phonemic representations of paronyms are identical, their lexical stress placement differs. For example: “φίλη” → [f`ili] and “φυλή” → [fil`i]. According to the intonational theory [24], lexical stress has an impact on the perceived tonal stress however it is not the only factor that affects the position of the prosodic pitch accent in an intonational phrase. There are stronger events, such as the phrase accent, which usually define a different tonal prominence in the word/phrase than the one imposed by the lexical stresses themselves. Phrase accent is defined from a series of linguistic attributes, such as the phrase type (e.g. affirmative, negative, question, wh-question etc). As a result, written lexical stress does not always produce tonal rises; falls are also common and this depends on both linguistic and para-linguistic phenomena, as well as the common ground between collocutors. This normal inconsistency between the written lexical stress and the acoustic pitch accent alignment leads to acoustic misunderstandings between words with identical phonemic representation but different lexical stress placement.

Though paronym perception accuracy has not been studied in depth, there are works that study the intelligibility of segments in both natural and synthetic voices, under normal or noisy conditions. In [5], the error rate in phoneme discrimination was measured to be 2.4% in natural speech, which was raised to 9.0% when using synthetic speech (Festival TTS, diphone-based voice). Most of the errors occurred in the final consonant. In that work it was also shown that in reverberation environments the phoneme discrimination is even lower. These values are significant higher than older works which showed that phoneme error rate in intelligibility tests using rule-based synthetic voices was ranging from 12.92% to 73.97% [6]. Finally in [7], it was shown that the initial segments of CV and VC syllables are harder to identify than the final segments, independently of the noise level.

The aims of the study presented here were: i. to categorize the types of errors the blind and sighted students made when rendering paronym words ii. to evaluate blind and sighted students' performances separately (within groups) and iii. to compare blind and sighted students' auditory discriminations (between groups).

2 Method

Sixty (60) university students took part in this study; thirty (30) of them had severe visual impairments (21 totally blind and 9 partially sighted) and the other thirty (30) students were sighted. The age range for the group of the sighted students was from 21 years to 24 years (mean= 22.27, SD= 1.081) and the age range for the group of the blind students was from 18 years to 35 years (mean= 26.23, SD= 4.804). They were all students of different schools in Greek universities and apart from blindness had no other additional disabilities.

Each student was interviewed and audiotaped in a 60 minute session. The experimental design comprised two parts: structured interviews and experiments. The interviews – apart from the participants' personal details - were focused on issues such as the experience of the students in using assistive technology mostly in the usage of screen readers. Each item in the questions was rated on a 4-point Likert-type scale, from 1 ("I use it a lot) to 4 ("I don't use it at all").

After the interview, every student was invited to listen to structured scripts and was asked to give back what he/she heard from the aural presentation of the scripts. The scripts were divided into two sets. The first one comprised 12 pairs of paronym words (Test 1) and the second one comprised 17 sentences in which were included also paronym words (Test 2). The auditory scripts were composed in conjunction with appropriate modifications and non-speech insertions. The paronym words as well as the structure of the sentences were chosen from the research team (authors) after many internal tests. The hypotheses of this study relied on the performances of the blind and sighted students when doing the tasks in recognizing the "paronyms" in specific auditory scripts.

These experimental sessions were carried out by utilizing a synthetic voice instead of a natural one. The purpose behind this decision was the fact that text-to-speech synthesizers have become essential tools for the accessibility of blinds and modern human machine interaction applications. Thus, the nature of the experiments presented in this work originates from real operational conditions. Before running the experiments, all students listened to examples of synthetic speech generated by DEMOSTHeNES language platform [8], [9], [10]. All subjects – one by one – were given 5 minutes to familiarize themselves by listening to some examples of synthetic speech which their content was different from the content of the stimulus material (words & sentences). After the familiarization stage the stimulus material was presented aurally by DEMOSTHeNES to the students and the latter were invited to render what they heard. The participants could stop the acoustic rendition simply by pressing a button in a device and write down in their own pace what they were listening from the headphones. The only restriction which took place was the fact they could not reverse the acoustic rendition to elucidate a misunderstanding. There were approximately 5 seconds break between the acoustic representations of the pairs of the paronym words as well as between the sentences which included also paronym words.

The default synthetic voice that we used featured a well established corpus-based prosodic model [11] and the Mbrola synthesizer [12] with the Greek diphone database gr2 [8]. The prosodic baselines used were: pitch=110Hz, speed=140 words per minute and volume=100.

The whole procedure was driven by the researchers, and, in the end, all the students' answers were audiotaped, transcribed, organized, reviewed for errors, and analyzed using SPSS 14.0.

Paronym words are characterized as linguistic structures which their acoustic representation bears a great resemblance between them and as such they might create misunderstandings or confusions. For this the categorisation of the students' errors regarding paronyms was based on phoneme errors. The phoneme error pattern used in this study was a synthesis of other similar patterns [13], [14] and comprised seven categories which referred to errors of phonological type. Phonological type errors are those that change the acoustic image of the word. These categories represent different cases such as omission of phonemes, addition of unneeded phonemes, wrong accentuation or different combinations of the above (see Table 1 for a list of all categories).

Table 1. Categories of Phonological-type errors (PTE)

Categories	Phonological-type errors
A	Accentuation (accent)
B	Phoneme substitution (first part of the word)
C	Phoneme substitution (middle and last part of the word)
D	Addition of a phoneme (first part of the word)
E	Addition of a phoneme (middle and last part of the word)
F	Combination of the following: i. Omission of more than one phoneme in the word ii. Addition of more than one phoneme in the word iii. Wrong rendering regarding accentuation combined with phoneme substitutions or omissions
G	Omission of the whole word or rendering of a different word

Category A referred to accentuation issues. Accent in Modern Greek is a short line which is written only above vowels indicating the way they are pronounced in the word.

The analysis below focuses on correlations between the groups of the participants (sighted and blind) and the categories of phonological-type errors (PTE) in two sets of psychoacoustic experiments: a. phonological-type errors in words (PTEW) and b. phonological-type errors in sentences (PTES).

3 Results

Preliminary analyses showed that according to our data a violation of the assumptions of normality took place; therefore, it was decided to use non-parametric tests. Also it is important to mention that the only mistakes the participants made were when they tried to render the paronym words. Hence it can be argued that the specific text-to-speech system did not constitute an external uncontrolled variable to threaten the validity of the experiments.

It was found from the interviews that the correlations between PTE and sex, age, degree of visual loss and frequency of usage of screen readers respectively were not significant.

It was also found that within the group of the participants with severe visual impairments there was not significant correlation between the number of PTEW and the number PTES. On the other hand, there was a medium positive correlation between the same variables within the group of the sighted students (**r$_{\mathbf{spearman}}$** (2-tailed) = 0.368, $p<0.05$).

The Mann-Whitney U Test was conducted to compare the number of phonological-type error (PTE) made by the two groups (sighted & blind). The analysis indicated that there was a significant difference between the performances of the blind and the sighted participants regarding paronym words either in pairs or in sentences (U= 259.00, Z= -2.835, $p<0.01$). More specifically the number of the blind participants' PTE was significantly less than the corresponding number of the sighted ones. A more detailed analysis revealed that the differences between the two groups of the participants were significantly different only regarding the PTEW (U= 237.00, Z= -3.172, $p<0.005$) and not the PTES.

It was also attempted a more discernible analysis regarding the categories mentioned in Table 1. More analytically, Table 2 provides descriptive data for the blind students (see below).

Table 2. Means of PTEW & PTES in all categories of PTE regarding blind students

	PTEW		PTES		PTE	
	Mean	SD	Mean	SD	Mean	SD
Category						
A	0.97	1.426	1.27	1.484	2.23	2.269
B	0.23	0.430	0.00		0.23	0.430
C	0.40	0.724	0.07	0.254	0.47	0.860
D	0.27	0.521	0.00		0.27	0.521
E	0.57	0.568	0.10	0.305	0.67	0.711
F	0.40	0.675	0.37	0.556	0.77	0.858
G	0.70	1.149	0.10	0.403	0.80	1.157
Total	3.53	2.649	1.90	1.807	**5.43**	3.579

Table 3 (see below) also provides descriptive data for the sighted students' performances. The two tables indicated that blind students performed more accurately in recognizing paronym words in both tests (M=5.43) compare with the performances of their sighted peers (M=8.00).

Table 3. Means of PTEW & PTES in all categories of PTE regarding sighted students

	PTEW		PTES		PTE	
	Mean	SD	Mean	SD	Mean	SD
Category						
A	0.83	1.053	0.93	0.944	1.77	1.406
B	0.13	0.346	0.03	0.183	0.17	0.379
C	0.50	0.572	0.03	0.183	0.53	0.629
D	0.47	0.507	0.13	0.346	0.60	0.563
E	0.73	0.691	0.00		0.73	0.691
F	1.20	0.961	0.47	0.730	1.67	1.295
G	1.63	1.650	0.90	0.759	2.53	1.925
Total	5.50	2.113	2.50	1.548	**8.00**	3.040

The Mann-Whitney U Test showed that the two groups differed statistically in categories D, F and G. That is to say, that the number of the PTE were statistically significant more on behalf of the sighted students rather than for the blind in the following categories: i. D (U= 305.00, Z= -2.501, p<0.05), ii. F (U= 268.500, Z= -2.798, p<0.01) and iii. G (U= 190.00, Z= -3.971, p<0.001).

Finally, distinguishing the PTE into PTEW & PTES the Mann-Whitney U Test indicated that regarding the test with the pairs of the paronym words (PTEW) the sighted participants made significant more errors than their blind peers in categories F (U= 228.00, Z= -3.541, p<0.001) and G (U= 289.50, Z= -2.538, p<0.05), whereas, regarding the test with the sentences (PTES) the sighted students made significant more errors than their blind peers in categories D (U= 390.00, Z= -2.053, p<0.05) and G (U= 183.00, Z= -4.617, p<0.001).

4 Discussion

In general, the blind students' performances were at higher level of distinction regarding the number of errors (PTEW & PTES) in both set of scripts. The fact that blind students surpassed their sighted peers it might be happened due to the experience of the former in listening to a big variety of pre-recorded study materials when developing their literacy skills [15], [16], [17].

As mentioned in the Results section, there was a medium positive correlation between the number of the PTEW and the number of the PTES regarding the performances of the sighted group. That is to say that medium level of inaccuracy regarding PTEW is also associated with medium level of inaccuracy regarding PTES. In other words the sighted students showed an internal coherence regarding their acoustic distinction towards paronyms. It seemed that the rendering of the paronyms caused a sort of confusion in the sighted ones rather than in the blind participants.

Additionally, the previous inference is supported by the results of the Mann-Whitney U Test. The analysis of the data revealed that there was statistically significant difference in the PTE scores of blind and sighted students, particularly in the test with the words (pairs of paronym words-PTEW). It seemed that the sighted students faced difficulties when they were asked to identify and render the pairs of the paronyms whereas they did not experience the same difficulties when they invited to do the same in the second experiment (paronym words embodied in sentences – PTES). Therefore, it might be argued that the second experiment with the sentences provided a frame of a logical content which eventually helped the sighted participants to figure out more easily the paronym words. Blind students in Greece especially those who attend university do not have the chance to read their books in Braille whereas their sighted peers do not have such problems. The Greek Ministry of Education does not provide students in tertiary education with books in Braille. As a result most university blind students are oriented to aural reading. Hence the acoustic images of the words – including the paronyms- might have been grouped in a more concrete and distinctive form in blind students' memory. The fact that blind students use more the aural way of reading is supported by the findings of many researchers by highlighting the phenomenon of the decline in using Braille [18], [19].

Also it is worth noting that the classification system of the phonological type errors adopted in this study, seemed to offer the opportunity for a detailed analysis regarding the different kind of errors. In particular, Tables 1 & 2 presented the means and standard deviations of the performances of both groups (pair of words and sentences). The dispersion of the errors in the word tests was larger in the categories F and G in favor of the sighted students which in turn entailed a significant differentiation between them in terms of statistics with respect to the two student groups (see also Results). Category F refers to a combination of errors and it might also reflect the confusion the sighted students experienced when performing the tasks. Category G was found to be common in PTEW and in PTES by which the two student groups differed significantly. As mentioned in the Method section (see Table 1), category G refers to omission of the paronym word or to a substitution of the paronym word with another. It may be argued that the sighed students used more intellectual links to render the words or the sentences they heard during the tasks and as a result the degree of the consolidation of the paronym words as intellectual schemata [20] was decreased.

5 Conclusions

As mentioned in the introduction there is little research on issues germane to paronym acoustic realization and as such it was felt that this study is important in helping to fill that gap. The results of this research may be considered of great interest because they are strongly linked with educational issues. The practical implications of intonational theories have an important impact on the auditory representations of documents which in turn may facilitate or hinder the education of blind and sighted students in schools [21], [22], [23].

The adopted categorization system regarding phonemic errors (see Table 1) seemed that served well this study and provided more insights about the nature and peculiarities of the errors made by blind and sighted students. Also these issues are strongly linked to spelling which in turn leads to the area of literacy and generally speaking to the curriculum [14].

Finally there is need for further research increasing the number of the participants and conducting experiments in different periods of their schooling. This data may lead to improvements of the synthetic speech regarding the distinctive rendering of special linguistic patterns such as paronym words in conjunction with the integration of phonemic error patterns.

Acknowledgements

The work described in this paper has been funded by the European Social Fund and National Resources under the HOMER project of the Competitiveness Programme: PENED, Greek General Secretariat of Research and Technology.

References

1. Raman, T.V.: An Audio View of (LA)TEX Documents, TUGboat. In: Proceedings of the 1992 Anuual Meeting vol. 13(3) pp. 372–379 (1992)
2. Hakulinen, J., Turunen, M., Raiha, K.: The Use of Prosodic Features to Help Users Extract Information from Structured Elements in Spoken Dialogue Systems. In: Proceedings of ESCA Tutorial and Research Workshop on Dialogue and Prosody, Eindhoven, The Netherlands, pp. 65–70 (1999)
3. Shriver, S., Black, A., Rosenfeld, R.: Audio Signals in Speech Interfaces. In: Proceedings of International Conference on Spoken Language Processing, Beijing, China (2000)
4. Xydas, G., Argyropoulos, V., Karakosta, T., Kouroupetroglou, G.: 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 2005) Las Vegas, Nevada SA, vol. 3, pp. 411–420 (2005)
5. Venkatagiri, S.H.: Segmental Intelligibility of Three Text-to-Speech Synthesis Methods in Reverberant Environments. Augmentative and Alternative Communication 20(3), 150–163 (2004)
6. Logan, J.S., Greene, B.G., Pisoni, D.B.: Segmental intelligibility of synthetic speech produced by rule, J. Acoust. Soc. Am. 86, 566–581 (1989)
7. Cutler, A., Weber, A., Smits, R., Cooper, N.: Patterns of English phoneme confusions by native and non-native listeners, J. Acoust. Soc. Am. 116(6), 3668–3678 (2004)
8. Xydas, G., Kouroupetroglou, G.: The DEMOSTHeNES Speech Composer. In: Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland, pp. 167–172 (2001)
9. Xydas, G., Kouroupetroglou, G.: Text-to-Speech Scripting Interface for Appropriate Vocalisation of e-Texts. In: proceedings of EUROSPEECH, Aalborg, Denmark pp. 2247–2250 (2001)

10. Xydas, G.: Machine Learning Models and Selection Methods for the Auditory Representation of Documents via Synthetic Speech with Enriched Prosody, PhD dissertation, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece (2006)
11. Xydas, G., Spiliotopoulos, D., Kouroupetroglou, G.: Modeling Prosodic Structures in Linguistically Enriched Environments. In: Lecture Notes in Artificial Intelligence, vol. 3206, pp. 521–528. Springer, Heidelberg (2004)
12. Dutoit, T.: An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, Boston, MA (1997)
13. Tindal, G.A., Marston, D.B.: Classroom-based Assessment: Evaluating Instructional Outcomes. Columbus, OH: Merrill (1990)
14. Argyropoulos, S.V., Martos, C.A.: Braille Literacy Skills: An Analysis of the Concept of Spelling. Journal of Visual Impairment & Blindness 100(11), 676–686 (2006)
15. Koenig, A.J., Holbrook, M.C.: Literacy Skills, In: Koening, A. J. Holbrook, M. C. (eds) Foundations of Education. vol. II, AFB Press, pp. 264–312 (2000)
16. Kapperman, G., Sticken, J.: Assistive Technology. In: Koening, A. J., Holbrook, M. C. (eds.) Foundations of Education. vol. II AFB Press, pp. 500–516 (2000)
17. McCall, S.: Accessing the curriculum. In: Arter, C., Mason, L. H., McCall, S., Stone, S. (eds.) Children with visual impairment in mainstream settings. London: David Fulton, pp. 29–40 (1999)
18. Miller, B.: Spelling Bees and Grammar Gorillas. The Braillle Monitor (November 1999) vol. 42(9) (Accessed August 8, 2004) Available online at http://www.nfb.org/slate/slss9904.htm
19. Argyropoulos, V., Martos, A., Leotskakos, B.: Blind students and spelling: An investigation into braille literacy skills. In: Proceedings of the ICEVI European Conference 2005: Education – aiming for Excellence, Chemnitz (2005) pp. 180–185 (2005)
20. Wadsworth, B.J.: Piaget's theory of cognitive and affective development (4th edn.). New York: Longman (1989)
21. Amato, S.: Standards for competence in Braille literacy skills in teacher preparation programs. Journal of Visual Impairment and Blindness 96(3), 143–154 (2002)
22. Arter, C., Layton, L.: Reading Preferences of Pupils with Visual Impairment. The. British journal of Visual Impairment 18(1), 41–44 (2000)
23. Fellenius, K.: Computer-based instruction for young braille readers in mainstream education-An evaluation study. Visual Impairment Research 1, 147–164 (1999)
24. Ladd, D.R.: Intonational phrasing: The case for recursive prosodic structure. Phonology 3, 311–340 (1986)