Multilingual Text-to-Speech Software Component for Dynamic Language Identification and Voice Switching

Paul FOGARASSY-NESZLY[1], Costin PRIBEANU[2]*

[1] BAUM Engineering,
8, Str. Traian Moşoiu, Arad 310175, Romania,
pf@baum.ro.

[2] National Institute for Research and Development in Informatics – ICI Bucharest,
8-10, Mareşal Averescu Blvd., Bucharest 011455, Romania,
pribeanu@ici.ro. (**corresponding author*)

Abstract: Text-to-speech synthesis is a critical feature of the applications developed for people with visual or reading disabilities. In the last years there has been an increasing interest in multilingual text-to-speech synthesis, which requires multilingual text analysis and language specific speech synthesis. In this case, the dynamic switching of the synthetic voice is needed in order to enhance the usability and user experience. This paper aims at presenting a software component for multilingual text-to-speech synthesis. The software has been developed and tested in four steps: alpha version (proof-of-concept), functional version (beta), commercial version, and implementation. The beta testing results showed a high accuracy of the language detection algorithms, which perform properly on texts having a variable degree of fragmentation. The commercial version has been then successfully implemented in two applications for visually impaired people: an automatic reading machine and a personal organizer for the blind and visually impaired users. Both implementations have been tested with users for usability and acceptance. The evaluation results showed that a device with this component is easier to use by visually impaired people.

Keywords: multilingual text-to-speech, dynamic language identification, voice switching, accessibility, assistive technologies, visually impaired users, usability.

1. Introduction

Text-to-speech (TTS) means converting a text written in a given language into speech signals. The TTS synthesis is typically done by using a synthetic voice that is available for one specific language.

The text-to-speech synthesis is a critical feature of the assistive technology applications that are developed for people with visual or reading disabilities (dyslectic or illiterate), in order to make the electronic documents accessible for them. There are many examples of such assistive technologies: the automatic reading machines, screen readers, portable computers with voice interface, and Braille displays.

In recent years, there has been a growing interest in the development of applications that are able to process texts written in two or more languages. Two examples of application areas that need a multilingual text-to-speech synthesis are: the education for all and the multi-cultural contexts (Carlson et al., 1990), Udvari-Solner & Thousand, 1996; Hasselbring & Glaser, 2000; Turunen & Hakulinen, 2001; Feraru et al., 2010; Bourlard et al., 2011; Tripathi & Shukla, 2014; Van Laere et al., 2016).

Multilingual (polyglot) text-to-speech synthesis requires dynamic language identification. The algorithms are different than those used for automatic language identification, since in this case the recognition process is performed asynchronous, on a continuous text stream.

If the text is written in more than one language, at each language change the user has to manually change the corresponding synthetic voice. Changing the synthetic voice during a lecture is uncomfortable for a user. Therefore, another requirement for a multilingual TTS is to automatically select and switch the available voice for the document language.

In this paper, a multilingual text-to-speech software component capable of performing both dynamic language identification and synthetic voice switching is presented. **The objective of this paper is to integrate the previous contributions (Fogarassy-Neszly & Gherhes, 2014; Fogarassy-Neszly et al., 2015) into a comprehensive framework.**

The software component has been developed in the framework of the innovation project iT2V.

As shown in Figure 1, the development cycle is featuring four steps: alpha version (proof-of-concept), functional version (beta), commercial version, and implementation in several applications.

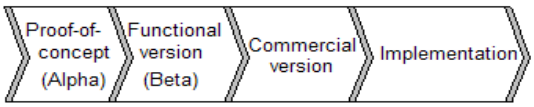

Figure 1. iT2V development cycle

The rest of the paper is structured as follows. Next section deals with the related work in the area of multilingual TTS and its applications. In section 3, the software component is presented. In section 4, the evaluation results are presented with a focus on the functional version. The paper ends with conclusions in section 5.

2. Related Work

2.1 Language identification

The statistical analysis of n-grams frequency is a widely used method for the language identification. A n-gram is a subsequence of n elements from a given sequence. In this work, a n-gram is a sequence of successive n-characters from a given text. According to the number of characters, we could have bigrams, trigrams, 4-grams, etc. Some examples of n-grams for a Romanian word are given in Table 1.

Table 1. Examples of n-grams

n-grams	analiză
2-grams	_a , an , na , al , li , iz , ză, ă_
3-grams	_an , ana , nal , ali , liz , iză , ză_
4-grams	_ana , anal , nali , aliz , liză iză_

The basic idea is that for each language there are n-grams that occur more often than others. In other words, the language could be identified based on the n-grams frequency. According to Canvar & Trenkle (1994), the use of tri-grams leads to the best results.

Language identification applications require a training phase for the acquisition of a relevant language corpus that will be further used during the language characterization phase. According to the studies of Dunning (1994) and Ljubesi et al. (2007), a corpus of 50 Kwords ensures a good accuracy.

The language corpus should be homogeneous and grammatically correct. The corpus quality is very important for the accuracy of the language identification.

The comparison between the analyzed spectrum and the reference one, can be done in different ways. The easiest criteria to use is the sum of absolute differences (Eq. 1), or the sum of the squared differences (Eq. 2).

$$A_L = \sum_{i=1}^{m} |f_{ai} - f_{Li}| \quad (1)$$

Where A_L is the frequency deviation for the language L, m is the number of n-grams in the analyzed text, f_{ai} is the frequency of the i-th n-gram in the analyzed text, and F_{Li} is the frequency of i-th n-gram in the frequency spectrum for language L.

$$A_L = \sum_{i=1}^{m} (f_{ai} - f_{Li})^2 \quad (2)$$

More sophisticated criteria for estimation of A_L deviation take into account through a weight coefficient the higher or lower probability (even zero) for certain n-grams in a certain language. In this case, the deviation is calculated according to Eq. 3.

$$A_L = \sum_{i=1}^{m} k_{Li} (f_{ai} - f_{Li})^2 \quad (3)$$

Where k_{Li} is the weight coefficient for i-th n-gram *in language L*; unlike in the equations (1) and (2), because k_{Li} could be positive or negative, A_L could be also positive or negative.

2.2 Multilingual TTS

Multilingual text-to-speech is a timely topic. Speech and voice-based technologies are a mean to overcome the access barriers to the web content for many people, including those with disabilities (Carlson et al., 1990). Text-to-speech can be used in education as a computer assisted learning tool (Mulyono & Vebriyanti, 2016). Text-to-speech is also a way to enhance the user experience with the computer technology since are offering an alternative modality to access the information.

Recent research in computer-based learning shows an interest to use TTS in schools with linguistic diversity. As Van Laere et al. (2016) pointed out, some pupils learning in a different language than the language spoken home are interested to use text-to-speech to improve understanding.

Many approaches for a multilingual TTS exist that differ as regards the solutions adopted for the text analysis and the speech synthesis Traber et al., 1999; Turunen & Hakulinen, 2001; Steinberger et al., 2006; Romsdorfer & Pfister, 2007; Shiga & Kawai, 2012; Chen et al., 2014; Ramani et al., 2014).

Most of the existing approaches are mainly targeted to the speech synthesis issues, aiming to synthesize the speech into a single speaker's voice (Bourlard et al., 2011; Shiga & Kawai, 2012; Chen et al., 2014; Ramani et al., 2014). There are relatively few approaches that focus on the voice switching issue.

Traber et al. (1999) distinguished between four categories of text-to-speech as regards the multilingual capabilities: monolingual, simple multilingual, mixed lingual with pre-defined language, and polyglot with language detection. In the case of monolingual TTS, the foreign words are rendered with the available voice. In the second case, the language change is accompanied by voice switching. In the third case, the system detects the foreign words, then adapts the pronunciation and intonation. In the last case, the TTS detects the language by using a multilingual text analysis and then generates the utterances by using phonetic and intonation models.

Romsdorfer & Pfister (2007) distinguished between a multilingual TTS synthesis based on manual language selection and a polyglot TTS synthesis that analyzes parts of text written in different languages.

3. The Software Component

The iT2V software component is featuring multilingual text analysis, automatic language detection and automatic language switching. In this respect, the software component plays the role of a voice independent layer between the application and the synthesis process. The language identification algorithms are based on the trigrams frequency comparison for a given text.

3.1 Proof of concept – alpha version

The goal of the alpha version was to test the language identification algorithms. In this case, the text was written in one language. The language identification method is based on the tri-grams frequency. In Figure 2, the user interface for the alpha version is presented (Fogarassy-Neszly & Gherhes, 2014).

The left part is used to enter the training text for each language (mini-corpus). The upper mid part is used to specify the candidate languages. The lower-mid part is for the display of tri-grams frequency. The right part is used to enter the text for testing and to display the results (scores for each candidate language).

3.2 Functional (beta) version

A preliminary functional version (beta01) has been developed that has three main modules: language configuration, training and dynamic recognition testing. The user interface enables the manipulation of two parameters: the look-

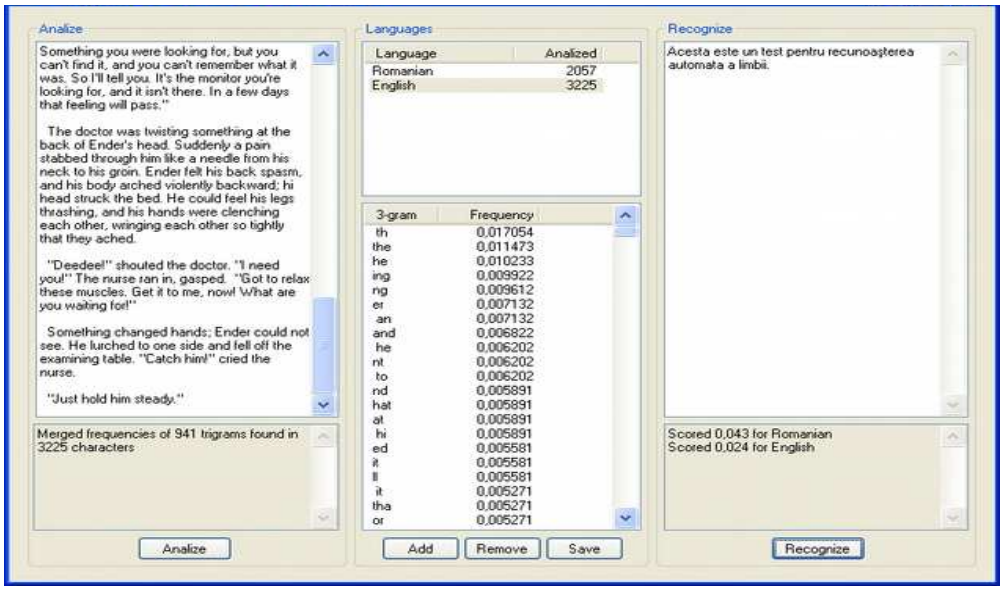

Figure 2. User interface for testing the statistical analysis algorithms

ahead (LA) specifying the number of words considered in the text analysis and the inertia (I) at language change.

3.2.1 Configuration module

The configuration module enables selecting the candidate languages (a minimum two and maximum of four languages), the desired synthetic voices, and to launch the other two modules. For each language a voice can be selected from a list of available voices.

3.2.2 Training module

The training module performs the language analysis for a given language corpus. The result is a set of tri-gram frequencies that could be further saved in a language file using an internal XML format (.lang).

Figure 3 presents an example of language analysis for the Romanian corpus, which resulted in 4.885 unique trigrams. These could be merged ("Merge results" button) with the existing data for Romanian language and eventually saved ("Save file" button).

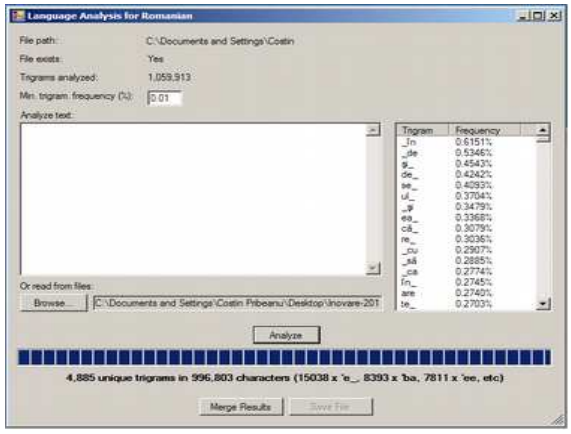

Figure 3. User interface of the training module

The user can load the training text directly in the „Analyze text" window or via a document file (.txt, .doc or .rtf).

The Beta02 version has been trained with corpora for the following languages: Romanian, English, Hungarian, French, and German. The number of tri-grams varied between 4.000 and 5.000.

3.2.3 Dynamic language identification module

The main module of the functional version enables the testing of language identification algorithms.

The user interface is presented in Figure 4. In this example, three candidate languages have been specified (Romanian, Hungarian, and English).

The user can introduce the text in the text box and then press the *Dynamic Recognition* button. After the text analysis, each piece of text is highlighted with the corresponding language color. This way, it is easier to detect the errors and assess the accuracy.

The user can also test the TTS and the voice switching by pressing the *Speak* button. At any time, is possible to cancel the text analysis or the vocal rendering by pressing the *Stop* button.

This module enables the manipulation of the two parameters: look-ahead (LA) specifying the number of words that are considered in the text analysis and the inertia (I) at language change.

The manipulation of these parameters is a very useful feature since it makes possible to fine tune the language detection algorithms. The first parameter influences the results, since the method is statistical and the precision depends on the size of the analyzed sample. It also influences the response time. In this respect, the optimal value is the smallest value for which the precision is acceptable.

The inertia parameter represents the degree to which the program delays the voice switching at language change. This parameter affects the user experience with a given application. If there is only one word in another language (for example "weekend" in a Romanian piece of text), it doesn't make sense to switch the voice.

After testing the beta01 version, it has been considered that switching the voice in the middle of a sentence is an unacceptable user experience problem.

Therefore, an improved functional version (beta02) has been developed and tested. The main improvement consists in voice switching at sentence level only.

3.3 Commercial version

The Speech Application Programming Interface (SAPI) used for the development of iT2V helps the developer to realize the speech synthesis through a standard interface.

The speech engine module developed by BAUM Engineering developers is according to Microsoft specifications, therefore this can be used through standard function calls. This is useful, because iT2V use other speech engine modules; these modules are not necessary obeying SAPI specifications.

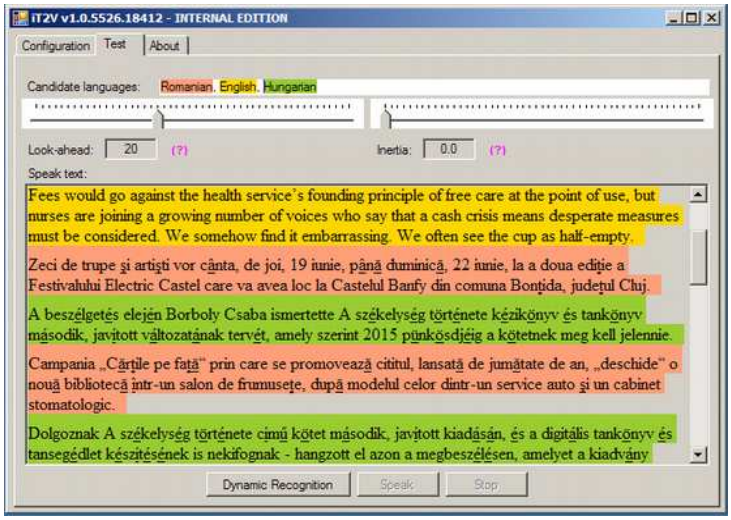

Figure 4. User interface of the dynamic language identification module

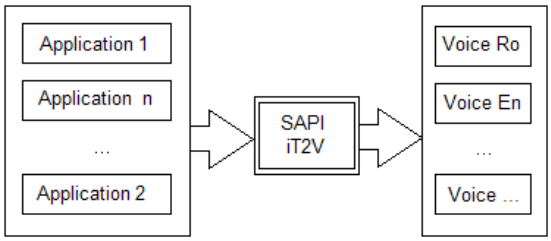

Figure 5. SAPI iT2V

The Speech Application Programming Interface SAPI-iT2V could be seen as an intermediate module between custom applications and speech engine modules of each supported language. This is why the iT2V component is presented by the system as a special voice, which should identify the language of the given text and select the right speech engine accordingly.

3.4 Implementations

The iT2V component has been successfully implemented in two assistive applications produced by Baum Engineering: the reading device POET, and the digital agenda Pronto. These applications are briefly presented in the next subsections.

3.4.1 Reading machine POET

The first implementation has been done on the automatic reading machine POET (trade mark of BAUM Retec AG) which is integrating a computer and a scanner (including the text recognition software with italics adjustment) in one single device. The device is simple and easy to use. It has two buttons (start and stop) and two control knobs to adjust the volume and the reading speed.

Figure 6 presents the reading device POET Compact 2 which user can easily employ in reading periodicals, books, or magazines. For Braille literates and deaf and blind users a Braille display can be connected to the Poet.

Figure 6. TTS reading device POET Compact 2

The document can be written on one or several columns. The text could be saved and then used for a new lecture or transferred to another device, such as an MP3 player. As Braille tables are language dependent, prior language detection is necessary before producing Braille output.

The model Compact 2+ has three buttons and a menu keypad and has been designed for more advanced users who prefer more features and full control over the functionality.

3.4.2 Digital organizer Pronto

Another implementation has been done on the organizer for blind people Pronto (trade mark of BAUM Retec AG).

Pronto is a portable organizer for visually impaired people, which runs under the operating system Window CE. The main functions are: taking notes with a text editor,

agenda & organizer, playing music, reading books in text or Braille format, Internet access.

Figure 7 presents the digital agenda Pronto. The device includes a Braille display and speech output.

Figure 7. Digital agenda Pronto

Pronto supports various applications, like a regular PDA (personal digital assistant). As such is more complex than other assistive technologies.

4. Evaluation

4.1 First functional version

The beta01 version has been tested with two, three, and four candidate languages. The LA and I parameters have been manipulated in order to identify which is their optimal range (Pribeanu & Fogarassy-Neszly, 2014).

The following measures have been collected: number of candidate languages (NCL), the number of language switching (NLS) the values of the look-ahead and inertia parameters, and the number of sentences for which the language is correctly detected. This measure serves to compute the effectiveness of language switching (EFS) as reported to the total number of sentences in the text.

The testing results revealed an acceptable accuracy for the LA parameter in the range 10-11 and the inertia parameter in the range 1.9-2.0. However, an important usability issue has been identified: switching the voice in the middle of a sentence which is not acceptable for the user.

4.2 The improved functional version

In order to enable comparison with the previous version, the same text has been used, but with a variable degree of fragmentation as regards the language change, respectively: low (NLS=3), moderate (NLS=10), and high (NLS=21). The first text has a distinct paragraph for each language (i.e., only 3 language switching).

The text contains sentences from newspapers (Adevărul, Times, Le Monde, and Uj Kelet). The text used for each language refers to at least two different domains. The text used for testing has a number of 8 sentences / 173 words in Romanian language, 10 sentences / 207 words in English, 9 sentences / 217 words in French, and 10 sentences / 161 words in Hungarian.

The results demonstrated an acceptable level of accuracy which has been higher for text with a low degree of fragmentation. The results were optimal for a look-ahead parameter in the range 25-30 and the inertia parameter in the range 0.0-0.3.

The results confirmed the improvements in the language detection algorithms. Table 2 presents the synthesis of results for the two beta versions (Fogarassy-Neszly et al., 2015).

Table 2. Synthesis of results for beta testing

Beta	NCL	NLS	LA	I	EFS
01	4	3	10	1.9	81.25%
	3	2	10	2.0	81.25%
	2	1	10	2.0	87.50%
02	4	3	30-40	1.0	97.30%
	4	10	20-30	0.0	91.89%
	4	21	20-40	0.3	91.89%
	3	10	25-30	0.0	96.55%
	3	15	25-30	0.0-0.3	96.55%
	2	11	24-30	0.0-0.2	94.44%

The degree of text fragmentation is given by NLS (number of language switching). The effectiveness of language switching (EFS) is higher for the text with low fragmentation and lower for the texts with moderate and high fragmentation. Also, the accuracy depends on the number of candidate languages.

4.3 Implementations

The evaluation has been done for the two assistive technologies previously presented: POET and Pronto. First, a usability inspection has been carried on by three experts. Then, each implementation has been tested with visually impaired users for usability and acceptance. After testing, participants were asked to answer a questionnaire as regards the factors that are influencing the technology acceptance (ease of use, usefulness, and enjoyment).

The applications have been tested with seven people (six men and a woman). The mean age of participants was 40.3 years (SD=9.23), with a minimum of 23 and a maximum of 48 years. The evaluation session took place in Arad, at the Local Branch of Romanian Association of the Blind People.

Except for one participant (university student), all are retired for medical reasons. The visual disability degree is severe (first degree). All participants have graduated a high school.

The participants have a variable degree of familiarity with the assistive technologies. The most commonly-used information and communication technologies are: mobile phone, computer, tablet and scanner + OCR. The most commonly-used applications are: Internet browsers, Skype, and Facebook. The most frequent goals are related to information, lecture, entertainment (games), and socialization. More details regarding the evaluation results could be found in Fogarassy et al. (2016).

The evaluation has been focused on the implementation of the iT2V component and not on the device itself. In this respect, no usability problem has been identified. In both cases, the accuracy of voice switching was excellent.

The users appreciated that the component saves time and make the use of these devices much simpler. The answers to the questionnaires show a difference between the ease of use of the two devices. Pronto has a more rich functionality which makes it more difficult to use. Nevertheless, in both cases the users have been satisfied and expressed the intention to use these technologies having iT2V.

5. Conclusion and Future Work

Multilingual text-to-speech brings many new challenges and opportunities to satisfy the needs of a large diversity of users. For people with visual disabilities, TTS is a critical feature.

In this paper a multilingual text-to-speech software component has been presented. The evaluation results of the beta version show that the fragmentation of a piece of text, as regards the language change, is an important parameter for the language detection algorithms.

The evaluation results of two implementations show that iT2V is usable, useful and enjoyable. A device with iT2V is easier to use. Since the users are visually impaired, manually changing the voice by following the guidance of an audio menu is difficult to use and time-consuming.

In the near future, the language recognition will be implemented in COBRA, a screen reader software. First of all, this will be very useful for the Internet browsing since beside pages in local language, pages in English could be accessed. Second, the implementation in a screen reader will be useful in the countries with two or more official languages or in multicultural contexts. Third, this facility will be useful for polyglot users interested in switching between documents written in different languages.

An unforeseen application of the language recognition is the Braille output according to a specific language Braille table. As the text to Braille translation depends on the language, without a specific table selection this job could be never automatically performed.

Acknowledgement

This work has been supported by the iT2V project (29DPST/2013), funded by UEFISCDI under the PNCDI II Innovation Program.

REFERENCES

1. BOURLARD, H., DINES, J., MAGIMAI-DOSS, M., GARNER, P. N., IMSENG, D., MOTLICEK, P., VALENTE, F. **Current Trends in Multilingual Speech Processing.** Sadhana, 36(5), 2011, 885-915.
2. CARLSON, R., GRANSTROM, B., HELGASON, P., JENSEN, P., TRAINSSON, H. **An Icelanding Text-To-Speech System For The Disabled.** STL-QPRS 31(4), 1990, 55-56.
3. CAVNAR, W., TRENKLE, J. **N-Gram-Based Text Categorization.** Proceedings of SDAIR-94, 1994, 161-176.
4. CHEN, C.P., HUANG, Y.C., WU, C.H. & LEE K. D. **Polyglot Speech Synthesis Based on Cross-lingual Frame Selection using Auditory and Articulatory Features.** Proceedings of IEEE/ACM TASLP 22 (10), 2014, 1558-1570.

5. DUNNING, T. **Statistical Identification of Language**. Technical Report MCCS 94-273, New Mexico State University, 1994.

6. FERARU, S. M., TEODORESCU, H. N., ZBANCIOC, M. D. **SroL – Web-based Resources for Languages and Language Technology e-Learning**. International Journal of Computers, Communications & Control, 5(3), 2010, 301-313.

7. FOGARASSY-NESZLY, P., GHERHES, V. **Applications for Dynamic Language Identification**. Proceedings of RoCHI 2014, Popovici D., M. & Iordache D.D. (Eds.), Constanta, 4-5 Sept., 2014, 51-54.

8. FOGARASSY-NESZLY, P., ZINVELIU, Z., PRIBEANU, C. **A Software Component for Polyglot Text-to-Speech Synthesis: User Interface and Beta Testing Results**. Proceedings of RoCHI 2015, Dardala, M., Rebedea, T.E. (Eds.), Bucharest, 24-25 Sept., 2015, 145-148.

9. FOGARASSY-NESZLY, P., PATRU A, IORDACHE D.D., PRIBEANU C. **Implementation of a Polyglot Text-to-Speech Synthesis in Two Assistive Technologies**. Proceedings of RoCHI 2016, Iftene, A., Vanderdonckt, J. (Eds.), Iasi, 8-9 September, 2016, in press.

10. HASSELBRING, T.S., & GLASER, C.H.W. **Use of Computer Technology to Help Students with Special Needs**. The Future of Children, 2000, 102-122.

11. LAERE, E. VAN, ROSIERS, K., VAN AVERMAET, P., SLEMBROUCK, S., & VAN BRAAK, J. **What Can Technology Offer to Linguistically Diverse Classrooms? Using Multilingual Content in a Computer-based Learning Environment for Primary Education**. Journal of Multilingual and Multicultural Development. 2016.

12. LJUBEŠIC, N., MIKELIC, N. & BORAS, D. **Language Identification: How to Distinguish Similar Languages**. Proceedings of the 29th International Conference on Information Technology Interfaces, 2007, 541–546.

13. MULYONO, H., VEBRIYANTI, D. N. **Developing Native-Like Listening Comprehension Materials Perceptions of a Digital Approach**. Journal of ELT Research, 1(1), 2016, 1-20

14. PRIBEANU, C., FOGARASSY-NESZLY, P. **Beta Testing of a Dynamic Language Identification Software Component - Preliminary Results**. Revista Romana de Interactiune Om-Calculator 7(3), 2014, 259-272.

15. RAMANI, B., ACTLIN JEEVA, M.P., VIJAYALAKSMI, P., NAGARAJAN, T. **Cross-lingual Voice Conversion-based Polyglot Speech Synthesizer for Indian Languages**. Proceedings of INTERSPEECH, 2014, 775-779.

16. ROMSDORFER, H., PFISTER, B. **Text Analysis and Language Identification for Polyglot Text-to-Speech Synthesis**. Speech Communication 49, 2007, 697-724.

17. SHIGA, Y. & KAWAI, H. Multilingual **Speech Synthesis System**. Journal of the National Institute of Information and Communication Technology 59 (3/4), 2012, 21-28.

18. STEINBERGER, R., POULIQUEN, B., WIDIGER, A., IGNAT, C., ERJAVEC, T., TUFIS, D., VARGA, D. **The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages**. Proc. LREC'2006, Genoa, Italy, 2006, 2143- 2146.

19. TRABER, C., HUBER, K., NEDIR, K., PFISTER, B., KELLER, E., & ZELLNER, B. **From Multilingual to Polyglot Speech Synthesis**. Proceedings of EUROSPEECH, Budapest, Hungary, 1999, 835–838.

20. TRIPATHI, M., & SHUKLA, A. **Use of Assistive Technologies in Academic Libraries: A Survey**. Assistive Technology, 26 (2), 2014, 105-118.

21. TURUNEN, M., & HAKULINEN, J. **Mailman-a Multilingual Speech-only e-mail Client based on an Adaptive Speech Application Framework**. Proceedings of Workshop on Multi Lingual Speech Communication - MSC 2000, 2000, 7-12.

22. UDVARI-SOLNER, A., & THOUSAND, J.S. **Creating a Responsive Curriculum for Inclusive Schools**. Remedial and Special Education, 17(3), 1996, 182-191.