# A Software Component for Polyglot Text-to-Speech Synthesis: User Interface and Beta Testing Results

**Paul Fogarassy-Neszly, Zlatomir Zinveliu**
BAUM Engineering
Str. Traian Moşoiu nr. 8, 310175 Arad
pf@baum.ro, zz@baum.ro

**Costin Pribeanu**
ICI Bucureşti
Bd. Mareşal Averescu nr.8-10, Bucureşti
pribeanu@ici.ro

## ABSTRACT

Text-to-speech synthesis has many applications in the area of assistive technologies for visually impaired people. Some applications require multilingual text-to-speech synthesis. In this case, multilingual text analysis and voice switching are desirable. In this paper an improved functional version (beta) of a software component for polyglot text-to-speech synthesis is presented. Beta testing results are useful to reveal the optimal level of parameters and suggest new directions to improve the method and algorithms. Testing results show that the new version of the component is able to automatically detect the language with a reasonable accuracy from texts with a variable degree of fragmentation.

### Author Keywords

Multilingual speech synthesis, language recognition, text-to-speech synthesis, assistive technologies beta testing, usability, accessibility

### ACM Classification Keywords

D.2.2: Design tools and techniques. H5.2 User interfaces.

## INTRODUCTION

Many assistive technologies for visually impaired people are using text-to-speech (TTS). TTS means converting a text document into speech signals, typically by using voice fragments prerecorded by a native speaking person in the desired language [8]. If the text is written in another language then the user has to manually select a corresponding voice for that language. The TTS synthesis is mainly used by people with visual or reading disabilities (dyslectic or illiterate) in order to make accessible the electronic documents. Examples of assistive technologies using TTS synthesis are: screen readers, automatic reading machines, portable computers with voice interface, and Braille display.

There is an increasing interest in applications based on assistive technologies that are able to process texts written in two or more languages. There are many application areas that need polyglot text-to-speech, such as education for all and multi-cultural contexts, to name just two [7, 12, 13, 14].

In this case both a multilingual (polyglot) text-to-speech synthesis and voice switching are needed. This requires to analyze the text piece by piece, to detect the language for each fragment and then to select the voice available for that language. Many approaches for multilingual TTS exist [1] (see [3], [6], [7], [9], [10], and [11]) that differ with regard to the solutions adopted for text analysis and speech synthesis.

Traber et al. [10] classified speech synthesis as regarding multilinguality into four categories: monolingual, simple multilingual, mixed lingual with pre-defined language, and polyglot with language detection. In the first case, foreign words are rendered with the available voice. In the second case, language switching is accompanied by voice switching. In the third case, the synthesis process detects foreign words and adapts the pronunciation and intonation. The fourth category is able to detect the current language using multilingual text analysis and use phonetic and intonation models to generate utterances.

Romsdorfer & Pfister [7] made a more clear distinction between multilingual TTS synthesis that need manually language selection and polyglot TTS synthesis that analyze parts of text in different languages. In this case, language identification of the text is indispensable.

The main objective of the research project iT2V is to develop and implement a software component for automatic language identification and voice switching. The project is carried on in a consortium of three partners: BAUM Engineering, ETA Automatizari Industriale, and National Institute for Research and Development in Informatics – ICI Bucharest.

The development follows four steps: alpha version (proof-of-concept, functional version (beta), commercial version, and implementation in several applications. The lifecycle of iT2V is illustrated in Figure 1.

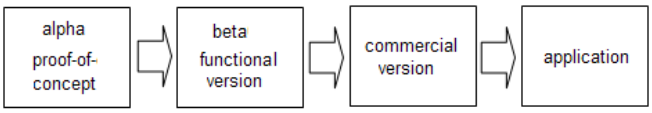


**Figure 1. iT2V lifecycle**

The alpha version was described in a previous work [4]. In this case, the text was written in one language and the goal was to test the language identification algorithms. In a recent paper a preliminary functional version (beta01) was presented. The evaluation results showed an acceptable accuracy when testing with four candidate languages [5]. However, switching the voice in the middle of a sentence is an important shortcoming for the user.

The objective of this paper is to present an improved functional version of the software component (beta02) and to present and discuss the evaluation results. The component has been tested with four, three and two candidate languages on texts with a variable degree of fragmentation. The results are analyzed against two

additional parameters: look-ahead (number of words considered in text analysis) and inertia at language switching.

The automatic language identification and voice switching are supported by a software component having the role of intermediate layer between the client application and the speech synthesis process. Unlike many other similar algorithms designed for language recognition, our approach is optimized for runtime; this is the main novelty of the proposed software component.

## THE SOFTWARE COMPONENT

### Functions

The software component supports a polyglot TTS and is able to perform multilingual text analysis, automatic language detection and automatic language switching. It plays the role of intermediate layer, voice independent, between the application and the synthesis process. Language identification is based on computing and comparing trigrams frequencies of a given text [2].

The functional version enables testing of implemented algorithms and analyzing the influence of various parameters on achieved results. It has three main modules that enable language configuration, training and dynamic recognition testing.

### Configuration module

The configuration module allowed user to select the candidate language (maximum 4 and minimum 2), the desired voices and to launch the language training and dynamic recognition functions. The user interface is presented in Figure 2.

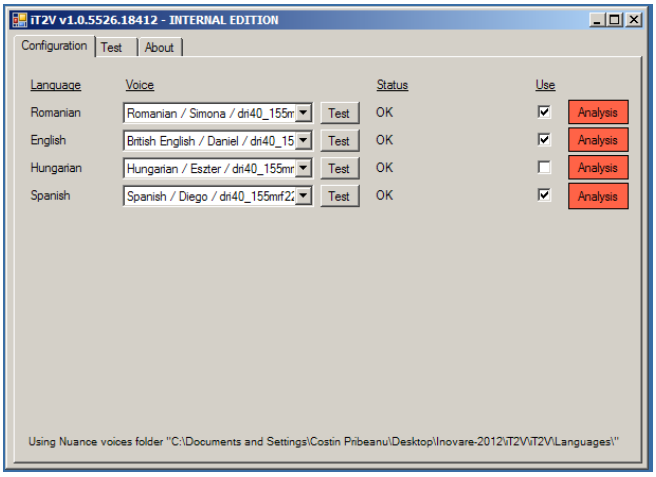


**Figure 2. User interface of the configuration module**

For each language a voice can be selected from a list of available voices. In the example above, the testing program is configured for 3 candidate languages.

### Training module

The language recognition component requires a training phase before use. The interface of the training module enables language analysis that results in tri-gram frequencies. These frequencies can be saved in a language characteristic file in an internal format (.lang). Figure 3 presents the user interface of the training module.

The text can be loaded directly in „Analyze text" window or via a document file (.txt, .doc or .rtf). It is possible to further extend the language file with other documents, if required.

Beta02 was trained with corpora for five languages: Romanian (Ro), English (En), Hungarian (Hu), French (Fr) and German (De). The number of trigrams frequency stored for each language varies from 4.000 to 5.000.

In Figure 2 an example of language analysis for the Romanian corpus is presented.

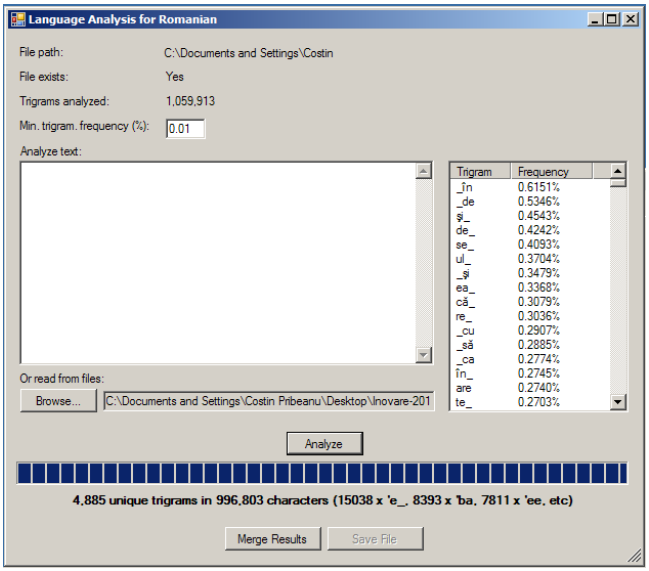


**Figure 3. User interface of the training module**

The analysis 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).

### Dynamic language recognition module

The main module of the software component enables testing the language recognition between any selections from available languages. The main improvement in beta02 consists in controlling for language switching at the sentence level only, avoiding the disturbing voice change in the middle of the phrase.

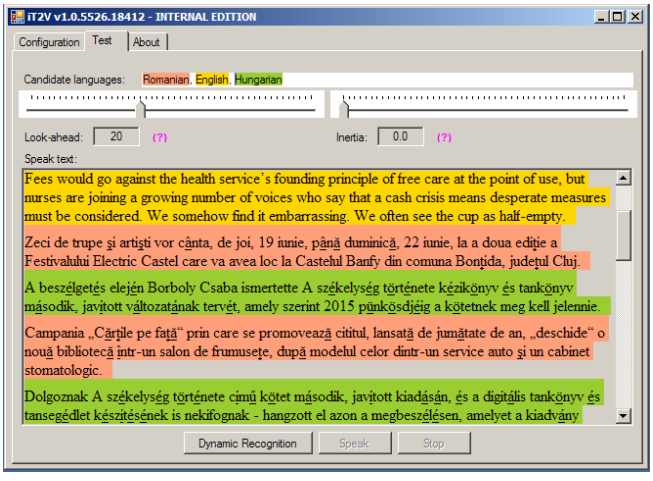


**Figure 4. Dynamic recognition language user interface**

The user interface is presented in Figure 4 where three candidate languages are specified (Romanian, English,

Hungarian, in this example) and marked with distinct colors.

After text analysis, the module highlights each piece of text with the corresponding language color. This makes it easier to detect errors and assess the precision.

There are two main parameters that could be varied within this module: look-ahead (LA) and inertia (I). The former is an integer specifying the number of words that are analyzed and the latter the inertia at language change, as a difference between computed criteria for candidate languages. The first parameter influences the results since the trigram method is statistical and the precision depends on the size of the analyzed sample. It also affects the response time so an optimal value is the smallest value for which the precision is satisfactory.

The second parameter represents the degree to which the program delays the language switching when the statistical criterion indicates another language for the text analyzed. This parameter affects the user experience. If there is just one word in another language (for example weekend in a Romanian text) it doesn't make sense to change the voice. Also, it is annoying for the user if the voice changes in the middle of a sentence. After testing the beta01 version, it was decided that for the target applications of iT2V is better to restrict language switching at sentence level.

## EVALUATION RESULTS

### Text used for testing

For testing with four candidate languages, three texts with the same content but different degree of fragmentation (number of language changes) were used: low (3), moderate (10), and high (21). The first text has a distinct paragraph for each language, hence there are only 3 language switching. For further testing with three and two languages the text content for the respective language was preserved.

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

An excerpt from the text with four languages (Ro, En, Fr, and De) and high degree of fragmentation is given bellow.

*Les ambulanciers ont dû remettre leurs téléphones portables lorsqu'ils ont vu le visage, qui aurait beaucoup maigri durant les 170 jours d'hospitalisation à Grenoble, après son grave accident de ski, en décembre, selon le journal. Bis zu den Wahlen 2016 in Somalia müssen unbedingt wirksame politische Lösungen für das Land gefunden und moderate Kräfte unter den Islamisten mit einbezogen werden. Il n'a pas parlé, mais il a communiqué avec les ambulanciers par des hochements de tête, durant les quelque 200 km du trajet. Pentru a convinge comisiile de examen că stăpânesc bine limba română, la nivel conversaţional, elevii au trebui să cunoască stilurile limbii române. They urged politicians to say how they plan to pay for a health service that faces a £30 billion funding black hole by the end of the decade. Patient leaders attacked the plan, saying that charges would deter the poor from seeking help, push people towards A&E and require cumbersome bureaucracy to collect. Stilul beletristic este specific romanelor şi încurajează folosirea figurilor de stil, a termenilor arhaici, regionali, jargou etc. Stilul juridico-administrativ apare atunci când vorbim de documente oficiale, iar stilul publicistic este caracteristic articolelor de presă. Es fehlt an Ausrüstung wie etwa Kampfhubschraubern, um die selbsternannten Gotteskrieger in Somalia zu bekämpfen. La chambre de commerce et d'industrie de Paris-Ile-de-France, qui représente 800 000 entreprises, s'inquiète du nombre croissant de jeunes envisageant de faire leur vie professionnelle à l'étranger.*

### Measures

Following measures were collected: number of candidate languages, number of language switching, look-ahead parameter, inertia parameter, and number of sentences for which the language was correctly detected. The last measure enables to compute the effectiveness of language switching (rate of success) as reported to the total number of sentences in the text.

### Procedure

Preliminary tests showed that the results are acceptable (effectiveness over 90%) when the parameters are varying in the range 10-40 for LA and 0.0-2.0 for I.

Testing has been carried on in six sessions. In the first three sessions the beta02 was configured for testing with four candidate languages (Ro, En, Fr, and De). LA parameter was varied with an increment of 5 and I parameter with an increment of 0.3. In the first session the text with low degree of fragmentation was used. In the second and third session the texts with moderate and high degree of fragmentation were used. Next two sessions used three candidate languages (Ro, En, and Hu) and the last session only two candidate languages (Ro and En).

### Results

The testing text with no fragmentation was also used for the evaluation of beta01 thus enabling comparison. A synthesis of testing results is presented in Table 1.

NCL is the number of candidate languages, NLS is the number of language switching, LA is the look-ahead parameter, I is the inertia parameter, and EFS is the effectiveness of language switching.

The beta testing results provide with useful information by showing the optimal level of parameters. The results for beta 02 are suggesting a look-ahead parameter in the range of 25-30 with inertia between 0.00 and 0.02.

**Table 1. Synthesis of results**

| 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% |

As it could be observed, the testing results for beta02 are better. The effectiveness of language switching (EFS) is 97.30% for the text with low fragmentation and 91.89% for the texts with moderate and high fragmentation. Since the text content is identical, this means that the text fragmentation is influencing the results.

## CONCLUSION AND FUTURE WORK

In this paper an improved functional version of a software component for polyglot text-to-speech synthesis was presented. The testing results of the beta02 functional version confirm the improvements of language detection algorithms.

Testing was done using texts with similar content but with different degree of fragmentation. The results suggest that text fragmentation is an important parameter for language detection algorithms.

In the next future the commercial version will be finalized and implemented in several applications.


### Acknowledgement

This work is supported by the IT2V research project (29DPST/2013), financed 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. (2011). Current trends in multilingual speech processing. *Sadhana*, *36*(5), 885-915.
2. Cavnar, W., and Trenkle, J. (1994). N-gram-based text categorization. Proc. 3rd Symp. on Document Analysis and Information Retrieval (SDAIR-94)
3. Chen, C. P., Huang, Y. C., Wu, C. H., Lee K. D. (2014). Polyglot speech synthesis based on cross-lingual frame selection using auditory and articulatory features. *IEEE/ACM TASLP 22 (10)*, 1558-1570.
4. Fogarassy-Neszly, P., Gherhes, V. (2014). Applications for dynamic language identification. Popovici D., M. & Iordache D. D. (Eds.) Proceedings RoCHI 2014, 51-54.
5. Pribeanu, C., Fogarassy-Neszly, P. (2014). Beta testing of a dynamic language identification software component - preliminary results. *Revista Romana de Interactiune Om-Calculator 7(3)*, 259-272, 2014.
6. Ramani, B., Actlin Jeeva, M.P., Vijayalaksmi, P., nagarajan, T. (2014). Cross-lingual voice conversion-based polyglot speech synthesizer for Indian languages. *Proceedings INTERSPEECH 2014*, 775-779.
7. Romsdorfer, H., Pfister, B. (2007). Text analysis and language identification for polyglot text-to-speech synthesis. *Speech Communication 49*, 697-724.
8. Shiga, Y. & Kawai, H. (2012). Multilingual speech synthesis system. *Journal of the National Institute of information and Communication Technology 59(3/4)*, 21-28.
9. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). *The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages*. In Proceedings of LREC'2006, 2142–2147, Genoa, Italy
10. Traber, C., Huber, K., Nedir, K., Pfister, B., Keller, E., Zellner, B. (1999). From multilingual to polyglot speech sinthesis. *Proceedigs of EUROSPEECH*, Budapest, Hungary, 835–838.
11. Traber, C., (1995). *SVOX: the implementation of a text-to-speech system for German*. Ph.D. thesis, No. 11064, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 7, ISBN 3 7281 2239 4).
12. Tripathi, M., & Shukla, A. (2014). Use of assistive technologies in academic libraries: A survey. *Assistive Technology*, *26*(2), 105-118.
13. Turunen, M., & Hakulinen, J. (2000, October). Mailman-a multilingual speech-only e-mail client based on an adaptive speech application framework. In *Proceedings of Workshop on Multi-Lingual Speech Communication - MSC 2000*, 7-12.
14. Udvari-Solner, A., & Thousand, J. S. (1996). Creating a responsive curriculum for inclusive schools. *Remedial and special education*, *17*(3), 182-191.


.