UNIUNEA EUROPEANĂ
GUVERNUL ROMÂNIEI
MINISTERUL MUNCII, FAMILIEI ŞI PROTECŢIEI SOCIALE
AMPOSDRU
Fondul Social European
POS DRU 2007-2013
Instrumente Structurale
2007-2013
MINISTERUL EDUCAŢIEI CERCETĂRII TINERETULUI ŞI SPORTULUI
OIPOSDRU
UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA

Investeşte în oameni!
FONDUL SOCIAL EUROPEAN
Programul Operaţional Sectorial Dezvoltarea Resurselor Umane 2007 – 2013

eng. **Adriana Cornelia STAN**

PhD THESIS

- ABSTRACT -

ROMANIAN HMM-BASED TEXT-TO-SPEECH SYNTHESIS WITH INTERACTIVE INTONATION OPTIMISATION

Scientific advisor,
Prof.dr.eng. Mircea GIURGIU

PhD Thesis Evaluation Committee:

PRESIDENT: - Prof.dr.eng. ***Dorin PETREUŞ*** - prodean, Faculty of Electronics,
Telecommunications and Information Technology,
Technical University of Cluj-Napoca;

MEMBERS: - Prof.dr.eng. ***Mircea GIURGIU***- scientific advisor,
Technical University of Cluj-Napoca;
- Prof.dr.eng. ***Corneliu BURILEANU*** - reviewer,
Politehnica University of Bucharest;
- Prof.dr.eng. ***Horia-Nicolai TEODORESCU,*** m.c.A.R. - reviewer,
“Gh.Asachi” Technical University of Iaşi;
- Prof.dr.eng. ***Aurel VLAICU*** - reviewer,
Technical University of Cluj-Napoca

2011

1 Introduction

1.1 Motivation

Speech synthesis has emerged as an important technology in the context of human-computer interaction. Although an intensive studied domain, its language dependency makes it less accessible for most of the languages. If for English, French, Spanish or German for example, the variety of choices starts from open-source user-configurable systems to high-quality proprietary commercial systems, this is not the case for Romanian. The lack of extended freely available resources makes it hard for the researchers to develop complete text-to-speech synthesis systems and design new language-dependent enhancements. The available Romanian synthesis systems are mainly commercial or based on outdated technologies such as formant synthesis or diphone concatenation.

There is also one other problem that is the main focus for the international research community, and that is the prosodic enhancement of the synthetic speech. Results of most of the speech synthesisers still have a monotone, unattractive flat intonation contour. This problem is usually solved by the use of fundamental frequency (F0) contour modelling and control of the parameters in a deterministic or statistical manner. Most of the F0 modelling or parametrisation techniques are based on extended speech corpora and manual annotation of the intonation. Some other solutions are language dependent methods, involving accent patterns or phrasing. Adaptation of these solutions to under-resourced languages is unfortunately unpractical and hard to achieve.

1.2 Objectives

Given the context presented before, the main objective of this thesis is the development of a new Romanian speech synthesiser, using the latest technology available. The system should also be able to allow for intonation adaptation. This challenge requires to address four specific objectives, as described below:

Objective 1: To develop a large high-quality speech corpus in Romanian and an associated word lexicon, which can support statistical training of the HMM models, but which can also be used for other speech-based applications.

Motivation: There are no Romanian speech corpora which can be used for statistical HMM training.

Objective 2: To create a Romanian text-to-speech system using state-of-the-art technologies, in the form of HMM-based parametric synthesis.

Motivation: The available Romanian TTS systems use either formant or concatenative synthesis. These types of synthesis methods have difficulties when trying to improve the naturalness or expressivity of the synthetic speech.

Objective 3: To design a new pitch modelling technique, which can be easily applied for intonation control.

Motivation: The existing pitch modelling techniques require extensive linguistic studies, and cannot provide a language-independent application.

Objective 4: To devise a method for interactive intonation optimisation of the synthetic speech.

Motivation: Even in state-of-the-art TTS systems, the expressivity of speech cannot be tuned by non-expert users.

2 Thesis Outline

The thesis is organised as follows:

Chapter 1 defines the motivation and the objectives of the thesis. It also outlines the thesis structure on a chapter by chapter basis.

Chapter 2 gives an overall view of speech synthesis methods with their respective advantages and disadvantages. A list of the available Romanian speech synthesiser is also presented. Chapter specific theoretical issues are presented on a chapter by chapter basis.

Chapter 3 introduces the preparation of the resources needed for a Romanian parametric speech synthesiser. After a brief introduction of the Romanian language characteristics, the chapter describes the tools and design procedures of both text and speech data. For text, the following issues are addressed: text corpus selection and preprocessing, phonetic transcription, accent positioning, syllabification and part-of-speech tagging. Speech resources include the recording of a high-quality speech corpus (about 4 hours) with the respective recording text selection and speech data segmentation. Two key features of the speech resources represent a list of semantically unpredictable sentences used for speech synthesis evaluation, and the preparation of a freely available online speech corpus (Romanian Speech Synthesis (RSS) corpus) which includes the recordings and several other information, such as HTS labels, accent positioning for the recorded text, or synthesised audio samples using RSS.

Chapter 4 presents the development of a Romanian HMM-based (Hidden Markov Model) speech synthesiser starting from the resources presented in chapter 3. A short theoretical overview of the HMM models and the HTS (HMM-based Speech Synthesis System) is presented. The preparation of HTS-compliant data is then described in terms of text annotation, decision tree clustering questions and segmentation and annotation of the training corpus. Apart from the novelty of a Romanian HTS system, the chapter introduces an evaluation of some language-independent configuration parameters. The results obtained are evaluated in a 3 section listening test: naturalness, speaker similarity and intelligibility.

Chapter 5 describes a novel approach to F0 parametrisation using the discrete cosine transform (DCT). The chapter starts by first analysing some of the most common F0 modelling techniques and their potential application in a system that uses no additional information, except from the text input and no complex phonological information. The DCT was chosen for its simplicity, language independency, high modelling capabilities even with a reduced number of features and the direct inverse transform useful in chapter 6. A superpositional model using the DCT is then proposed and evaluated in the context of both modelling and prediction of the F0 contour.

Chapter 6 uses the results of chapter 5 to define an interactive optimisation method using evolution strategies. The method uses the phrase level DCT coefficients of the F0 contour in a interactive CMA-ES (Covariance Matrix Adaptation - Evolution Strategy) algorithm. The basics of evolutionary computation are presented with a focus on evolution strategies and CMA-ES. Evaluation of the applicability scenario is performed. This includes the analysis of the initial standard deviation of the population, number of individuals per generation and dynamic expansion of the F0 contour. Results of a naturalness and expressivity listening test are analysed.

Chapter 7 summarises the main conclusions of the thesis and future development possibilities.

3 Discussion on Thesis Results and Future Work

3.1 Resource Development for a Romanian Parametric Speech Synthesiser

Resource development in a new language is an important step in creating any new speech processing system. The analysis of extended corpora can provide more accurate results. Text resources, as well as speech resources were gradually introduced within chapter 3. Although the resources cover a wide variety of aspects, they can only be viewed as a starting point for a more complex and elaborate source of information.

The text resources include a newspaper text corpus, simple letter-to-sound rules, accent positioning, syllabification and part-of-speech tagging. Text resources were not the main focus of the research and therefore each have identified problems. The text corpus contains around 4000 newspaper articles. Although the mass-media language is considered to be the reference for most of the speakers, it is not necessarily an optimal source for language studies. Literary works should also be included in such a resource. The phonetic transcriber written in Festival only included a minimal set of rules, which do not cover all of the rules described by phoneticians. Although, this can be argued against with the results of the intelligibility listening test.

A good resource for accent positioning is the DEX, but it is not practical to use an entire word database in a text processor. Even if Romanian does not have deterministic accent positioning rules, the accent could be derived using machine learning algorithms. The preliminary evaluation of the syllabification using MOP principle is only preliminary and its results cannot be taken for granted. A more extensive analysis in conjunction with the standard syllabification rules for Romanian should be performed. Part-of-speech tagging was also determined from an external source and cannot be fully controlled so far.

The developed lexicon includes accent positioning and phonetic transcription. Although, an extended and important resource, it should be modified so as to include more information, such as syllabification for example.

The speech resources developed in the context of this thesis are potentially one of the greatest contributions, considering the lack of such resources for Romanian. The design of the speech corpus makes it easy to use in many types of speech processing applications, such as automatic speech recognition, speech coding and of course speech synthesis. Its high-quality and sampling frequency are also an important feature. The inclusion of both newspaper and fairytale text in the recorded speech makes it more comprehensive. The entire speech corpus is freely available under the name of Romanian Speech Synthesis (RSS) database. A possible extension of this resource is naturally the recording of more speech data.

3.2 A High Sampling Frequency Romanian Parametric Text-to-Speech Synthesiser based on Markov Models

As it has been shown in chapter 4, Romanian language lacks proper open source synthesis systems for research. The HMM-based synthesis system along with the developed resources is an important addition to the research domain. Chapter 4 included the data preparation steps from text processing to speech segmentation and annotation. A first problem of the TTS system is the lack of an optimal text processor, with full text normalisation and POS tagging. Although it has not been proven, correct POS tagging could influence the result of the synthesiser.

The results of the listening test showed that the speech resources, configuration parameters and sampling frequency were appropriately selected. The evaluation of the system also included the evaluation of the amount of training data. It is commonly known that for speech synthesis, the larger the speech corpus, the better the results. However, an interesting development would be the selection of a minimum duration speech corpus with results comparable to the ones achieved herein. The listening test also showed that for Romanian, general-purpose designed semantically unpredictable sentences cannot determine significant differences between systems. Thus, a more complex method of evaluating

Romanian intelligibility should be designed.

3.3 A Language-Independent Intonation Modelling Technique

Chapter 5 was an evaluation of the parametrisation and prediction capabilities of the DCT transform for F0 contours. The proposed method makes a clear separation between the syllable and phrase levels of the fundamental frequency for F0 parametrisation. Each layer is individually modelled using a limited number of DCT coefficients. The statistical analysis of the DCT coefficients showed that as the order of the coefficient increases, the relative standard deviation decreases, which means less variability. It can therefore be concluded that by extending the number of DCT coefficients, no further major improvement can be achieved.

Each of the DCT coefficients is predicted separately using CART algorithms. The features used for the training vector are the ones available in the HTS label format, and thus no additional processing is required. CART algorithms are fast and efficient methods of estimation for low-complexity problems. As the results showed, their performance for high order coefficients is drastically reduced. This means that the analysis of some more advanced machine learning methods, such as neural networks or Markov models is needed. Also because of the separate estimation of the coefficients, some joint features can be overlooked. A joint estimation mechanism would probably enhance the prediction results.

The attribute selection provided the means for a complexity reduction of the problem, but it did not provide accurate correspondence between the DCT coefficients and the phonological features used in the feature vector. A more elaborate analysis of this correspondence should also be performed.

3.4 Optimising the F0 Contour with Interactive Non-Expert Feedback

Language-independent F0 contour optimisation is a very important aspect of the speech synthesis domain. The method and prototype system presented in chapter 6 can be easily adapted to any HMM-based synthesiser with minimum adjustment. Preliminary evaluations carried out proposed the setup parameters of such a system and have shown that the dynamic pitch expansion can be achieved even with a small number of individuals and generations.

As the results obtained in this preliminary research have achieved a high-level of intonational variation and user satisfaction, a web-based application of the interactive optimisation is under-way. The application would allow the user to select the entire utterance or just parts of it – i.e., phrases, words or even syllables – for the optimisation process to enhance. For a full prosodic optimisation, the duration of the utterance should be included in the interactive application as well.

One drawback to the solution is the lack of individual manipulation of each of the 7 DCT coefficients in the genome, unattainable in the context of the evolutionary algorithm chosen. However the coefficients' statistics showed that the average standard deviation is similar and thus the choice for the initial standard deviation does not alter the higher order coefficients.

An interesting development would be a user-adaptive speech synthesiser. Based on previous optimisation choices, the system could adapt in time to a certain prosodic realisation. Having set up the entire workflow, testing different types of fitness functions is also of great interest.

4 Thesis contributions

The main contributions of the thesis are organised in chapters 3,4,5 and 6 and can be summarised as follows, along with their chapter correspondence and published papers:

1. A 65,000 Romanian word lexicon with phonetic transcription and accent positioning

Phonetic transcription and accent positioning represent two key aspects of a text processing module for text-to-speech synthesis. The 65,000 word lexicon represents 4.7% of the total entries of the DEX online database. The phonetic transcription was performed using the standard phoneme set for Romanian, excluding allophones and rare case exception pronunciations. Simple initial letter-to-sound rules were written in Festival, and some other rules were added manually in the lexicon. Accent positioning was directly extracted from the DEX online database.

The lexicon is an important linguistic resource mainly because of its dimension and contents. To the best of the author's knowledge there are no available resources of this type. The correctness of the information within was tested through the use of the lexicon in the front-end training of the Romanian speech synthesiser.

This contribution is supported by the development of the following additional resources:

- A text corpus of 4506 short newspaper articles trawled between August 2009 and September 2009 from the online newspaper "Adevărul". It contains over 1,700,000 words, and the top 65,000 most frequent were used in the lexicon;
- A reduced set of Romanian letter-to-sound rules written in Festival format for the initial phonetic transcription of the lexicon;

2. The Romanian Speech Synthesis (RSS) corpus: A high-quality broad application Romanian speech resource

Starting from the requirements of a parametric HMM-based speech synthesiser, the development of an extended speech corpus was identified. The Romanian Speech Synthesis corpus has a duration of 4 hours and comprises the following data:

- Training set utterances - approx. 3.5 hours
  - 1493 random newspaper utterances
  - 983 diphone coverage utterances
  - 704 fairytale utterances - the short stories Povestea lui Stan Păţitul and Ivan Turbincă by Ion Creangă
- Testing set utterances - approx. 0.5 hours
  - 210 random newspaper utterances
  - 110 random fairytale utterances
  - 216 semantically unpredictable sentences

The recordings were performed at 96kHz, 24 bits per sample and downsampled at 48kHz using professional recording equipment. The entire corpus, along with ortographic and phonetic transcription, time-aligned HTS labels, and accent positioning are freely available at `www.romaniantts.com`, and represent the most extended Romanian speech corpus.

The corpus was tested through its use in the model training part of the Romanian HMM-based speech synthesiser and also in a simple unit selection concatenative system. The semantically unpredictable sentences were evaluated as part of the intelligibility section of the listening test. The fairytale utterances have been used for the adaptation of the baseline trained models, in order to achieve a more dynamic intonation of the output speech. Statistic analysis of the recorded text within the speech corpus show similarities to the statistical distributions of the Romanian language.

This contribution is supported by the development of the following additional resources:

- The development of a list of 216 Romanian semantically unpredictable sentences used in speech synthesis evaluation. To the best of the author's knowledge, this is the first resource of this sort;
- A basic Romanian text processor for the HTS format labeling of the speech corpus.

3. An evaluation of the configuration parameters for the HTS system

HMM-based statistical parametric speech synthesis has become one of the mainstream methods for speech synthesis. The HTS framework offers a large number of possibilities for the parameter tunning of the generic system. The evaluation of the configuration parameters included the frequency warping scale, spectral analysis method, cepstral order, sampling frequency and amount of training data. The first three were heuristically determined based on analysis-by-synthesis methods, while the last two are evaluated within the listening test for the Romanian HTS synthesiser.

The results showed that:

- there are no significant perceptual differences between the Bark and ERB frequency scales when using the vocoder for 48kHz input data;
- the data driven generalised logF0 was validated;
- the MGC performed better than the mel-cepstrum analysis method;
- the cepstral analysis order is dependent on the sampling frequency;
- the use of high-sampling frequency increases the quality of the output speech, but the differences between 32kHz and 48kHz are not significant;
- an increased dimension of the training speech corpus enhances the quality of the synthetic speech.

4. A Romanian HMM-based speech synthesiser

The developed TTS system uses HMM-based statistical parametric speech synthesis, which is the latest technology available for speech synthesis. Employing the text and speech resources developed priorly, and the established configuration parameters, a number of 5 distinct systems were trained. They differ by the amount and sampling frequency of the training data.

The systems have been evaluated by 54 listeners, in a Blizzard-style listening test comprising 3 sections: naturalness, speaker similarity and intelligibility and along with a minimal unit selection concatenative system and the original recordings. The results of the listening test showed an average 3.0 MOS score for all of the HTS systems built, and an average of 3.3 MOS score for the best evaluated one. Sampling frequency has influenced the speaker similarity, but not the naturalness, while the amount of the training data had an effect on both sections. The WER in the intelligibility section, for all the systems was below 10%.

All of the HTS systems outperformed the unit selection system. Additionally, they have the capability to adapt to a more dynamic intonation speech corpus, as proved by the adaptation to the fairytale speech subset.

An interactive demonstration of the Romanian HTS synthesiser is available at `www.romaniantts.com`.

This contribution is also supported by the following additional elements:

- A set of 179 Romanian phonetic decision tree questions for context clustering in the HTS system;
- A basic text processing tool using the Cereproc Development Framework with minimal text normalisation and which outputs HTS format labels;

5. A language-independent F0 modelling technique based on the discrete cosine transform

This contribution solves the F0 modelling as a part of the language-independence issue for text-to-speech systems. The method adheres to the superpositional principle of pitch by modelling the syllable and phrase level contours, and uses a discrete cosine transform parametrisation. Only the textual features available in the HTS labels, without any additional linguistic information, and the DCT coefficients of the F0 contour are used for pitch modelling and prediction.

F0 prediction was performed using independently trained classification and regression trees, for each of the DCT coefficients. The results revealed an average error of 15Hz per utterance, which is similar to other modelling techniques. Also, the listening test showed that the users did not consider the differences between the HTS generated F0 contour, and the DCT predicted one as perceivable.

This contribution is supported by the following additional analysis:

- Statistic evaluation of the of the DCT coefficients within the *rnd1* subset of the RSS database;
- Evaluation of the DCT coefficient prediction results using 3 CART algorithms: M5 rules, Linear Regression and Additive Regression;
- Objective and subjective evaluation of the F0 contour estimation from the tree-based prediction of the DCT coefficients.

6. A method for the application of interactive CMA-ES in intonation optimisation for speech synthesis

The interactive intonation optimisation method solves a complex problem related to the expressivity enhancement of the synthesised speech, according to a non-expert listener's subjectivity. The originality of the method consists in using no prosodic annotations of the text, no deterministic rules and no predefined speaking styles. CMA-ES is applied in an interactive manner to the DCT coefficients of the phrase level F0 contour generated by the Romanian HTS system.

The main parameters of the interactive CMA-ES are evaluated and include:

- initial standard deviation of the population used to control the naturalness of the speech output, by limiting the domain of the F0 values;
- population size used to minimise user fatigue while maintaining a sufficient number of different speech samples the user can opt for;
- dynamic expansion of pitch over a number of generations to determine the evolution of the pitch contour according to the user's choices.

These parameters are also evaluated in the interactive intonation optimisation prototype system. To the best of the author's knowledge, this is also the first application of an interactive CMA-ES algorithm.

7. A prototype interactive intonation optimisation system using CMA-ES and DCT parametrisation

The proposed interactive intonation optimisation method has been implemented in a prototype system. The system is language-independent and uses the developed Romanian HTS system and the interactive CMA-ES parameters determined before. Given the output of the baseline speech synthesiser, the user can opt for further enhancements of the intonation for the synthesised speech. Four new different speech samples derived from the original F0 contour are presented to the listener in a tournament like comparison method. Starting from the overall winner of one generation, the next 4 individuals are generated.

The results of the prototype system have been evaluated in a listening test comprising naturalness and expressivity sections. The individuals naturalness was evaluated with an average MOS score of 3.1, and all of the newly generated speech samples were considered to be more expressive than the original one. Thus proving that the prototype system is able to maintain a natural output speech, while enhancing its expressivity.

The contributions can be included in the general processing scheme of an HMM-based speech synthesis system according to Fig. 1 .

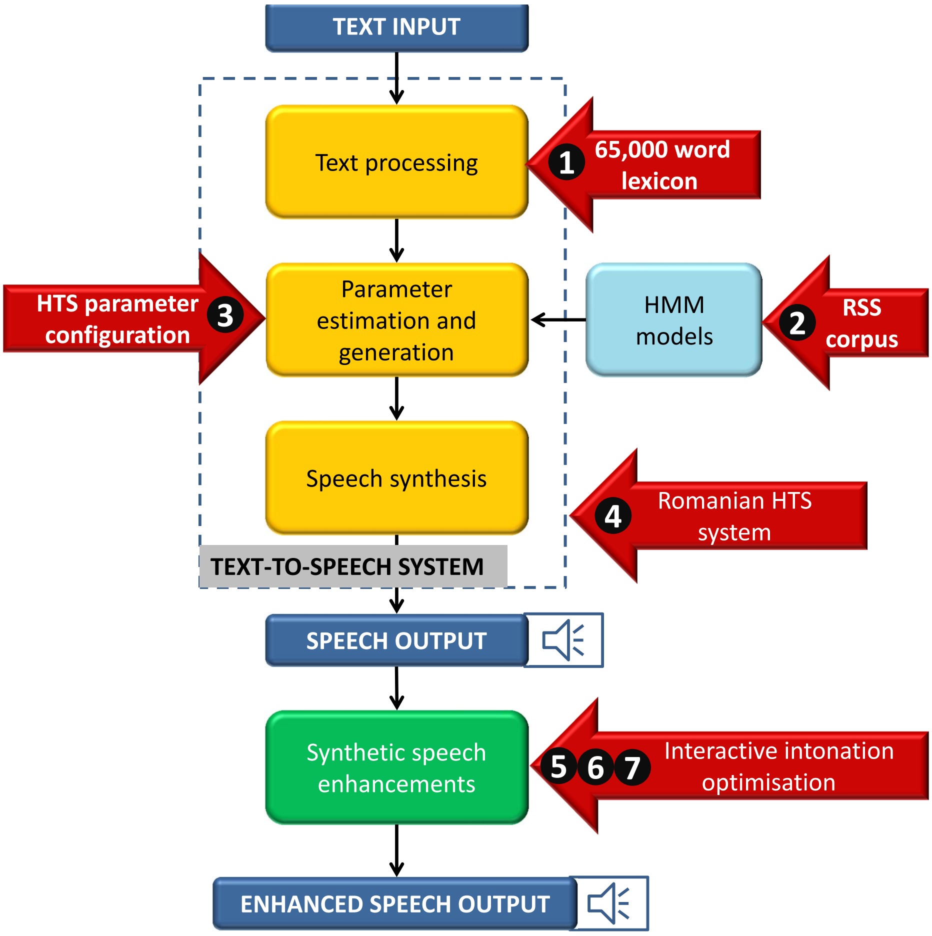

Figure 1: The application of the thesis contributions within the general processing scheme of an HMM-based speech synthesis system (marked with numbers from 1 to 7).

5 List of publications

Journals

1. **Adriana STAN**, Junichi YAMAGISHI, Simon KING, Matthew AYLETT, *The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate*, Speech Communication, vol 53, pg. 442-450, 2011.

Conference Proceedings

1. **Adriana STAN**, Florin-Claudiu POP, Marcel CREMENE, Mircea GIURGIU, Denis PALLEZ, *Interactive Intonation Optimisation Using CMA-ES and DCT Parametrisation of the F0 Contour for Speech Synthesis*, In Proceedings of the 5th Workshop on Nature Inspired Cooperative Strategies for Optimisation, in series Studies in Computational Intelligence, vol. 387, Springer, 2011.
2. **Adriana STAN**, Mircea GIURGIU, *A Superpositional Model Applied to F0 Parametrisation using DCT for Text-to-Speech Synthesis*, In Proceedings of the 6th Conference on Speech Technology and Human-Computer Dialogue, doi: 10.1109/SPED.2011.5940734, Braşov, România, 18-21 May 2011,
3. **Adriana STAN**, Mircea GIURGIU, *Romanian language statistics and resources for text-to-speech systems*, In Proceedings of the 9th Edition of the International Symposium on Electronics and Telecommunications, pg. 381-384, Timişoara, România, 11-12 November 2010.
4. **Adriana STAN**, *Linear Interpolation of Spectrotemporal Excitation Pattern Representations for Automatic Speech Recognition in the Presence of Noise*, In Proceedings of the 5th Conference on Speech Technology and Human-Computer Dialogue, pg. 199-206, Constanţa, România, 18-21 June 2009.

Scientific reports

1. **Adriana STAN**, *Raport de cercetare ştiinţifică 1: Elaborarea şi dezvoltarea unui sistem de sinteză text-vorbire în limba românăbazat pe modele Markov, independent de elementele de prozodie aferente textului*, May, 2010
2. **Adriana STAN**, *Raport de cercetare ştiinţifică 2: Elaborarea şi dezvoltarea unor metode deterministe de analiză şi control a prozodiei în limba română*, January, 2011
3. **Adriana STAN**, *Raport de cercetare ştiinţifică 3: Elaborarea şi dezvoltarea unor metode probabilistice de analiză şi control a prozodiei în limba română*, April, 2011

Acknowledgment

First of all I would like to thank my PhD advisor, prof.dr.eng. Mircea Giurgiu, for his continuous support and guidance.

Thanks to the Centre for Speech Technology Research of the University of Edinburgh, for a very welcoming and fruitful research visit, especially to dr. Simon King, dr. Junichi Yamagishi and Oliver Watts. Also to the fellow research assistants and PhD students with whom I've shared and discussed some ideas.

Also I would like to thank the Cereproc team, Matthew Aylett, Chris Pidcock and Graham Leary for the help with the text processor.

Thanks to Florin Pop, Marcel Cremene and Denis Pallez for the insight offered in the evolution strategies and the interactive intonation optimisation.

To Cătălin Francu, the author of the online DEX database and to the authors of the Romanian POS Tagger, dr. Doina Tătar, Ovidiu Sabou and Paul V. Borza.

*Adriana Stan was funded by the European Social Fund, project POSDRU/6/1.5/S/5.

*Parts of this work made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF – http://www.ecdf.ed.ac.uk). The ECDF is partially supported by the eDIKT initiative (http://www.edikt.org.uk).